THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR : Robert David Skeel

TITLE OF THESIS: Convergence of Multivalue Methods for Solving Ordinary Differential Equations

DEGREE FOR WHICH THESIS WAS PRESENTED: Doctor of Philosophy.

YEAR THIS DEGREE GRANTED: 1974

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to produce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

THE UNIVERSITY OF ALBERTA

CONVERGENCE OF MULTIVALUE METHODS FOR SOLVING ORDINARY DIFFERENTIAL EQUATIONS

by

ROBERT DAVID SKEEL

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTING SCIENCE

EDMONTON, ALBERTA

SPRING, 1974

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled CONVERGENCE OF MULTIVALUE METHODS FOR SOLVING ORDINARY DIFFERENTIAL EQUATIONS submitted by Robert David Skeel in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

[illegible]

[illegible]

[illegible]

[redacted]

[illegible]

ABSTRACT

Two variants of the multivalue method, the corrector only and the $P(EC)^M$ methods, for the numerical solution of the differential equation

$$y^{(p)} = f(t,y,\dots,y^{(p-q)})$$

are studied. It is shown that the numerical solution of a convergent corrector only method is also the solution of a stable multistep method. A useful technique for analyzing multivalue methods is discovered which makes it possible to formulate a constructive definition for the order of a method. When the stepsize is fixed, necessary and sufficient conditions for rth order q-convergence are given, and explicit expressions for the asymptotic behaviour of the global error are derived. When the stepsize is varied in a reasonable way, it is shown that the property of rth order q-convergence is preserved for $q = 1$ but not for $q > 1$. In the special case of a $(k+1)$-value Adams method, stability can be guaranteed by keeping the stepsize constant for at least $k-1$ steps after a change in stepsize.

ACKNOWLEDGEMENT

I wish to express my appreciation to Professor L.W. Jackson for his criticism and interest during the preparation of this thesis. I would also like to thank Professor S. Cabay for reading a draft of this thesis.

I am grateful to Mrs. M. Yiu for typing this manuscript. Finally I would like to thank the National Research Council of Canada for providing financial support during the past four years.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The objective of this thesis is to study multivalue methods for solving ordinary differential equations of the form

$$y^{(p)} = f(t,y,\ldots,y^{(p-q)}) \ .$$

Multivalue methods have been developed by Gear (1971) as generalizations of Nordsieck's method, which saves approximations of the higher order derivatives of the solution without directly evaluating these derivatives. In his original paper Nordsieck (1962) demonstrates the equivalence of his method to an Adams-Moulton method. Subsequently Osborne (1966) showed that the Nordsieck methods were equivalent to multistep methods.

Gear (1967), in extending the work of Descloux, noted that any multistep method can be written in many forms. To take a step requires a set of back values and derivatives, and this set defines an interpolating polynomial. There are many ways to represent this polynomial, and in particular, Nordsieck's method saves the higher order derivatives of the polynomial evaluated at the latest value of the independent variable t.

These saved values are denoted by the vector

$$\underline{a}_n = \mathrm{col}(y_n, hy_n', \ldots, h^{k+p-1} y_n^{(k+p-1)}/(k+p-1)!)$$

where h is the stepsize. This particular formulation of a (k+p)-value method is useful because it makes stepsize changing easy.

Gear (1971) has given the theory of multivalue methods considerable attention but some of the analysis remains incomplete. In particular, the classical result of "stability and consistency are equivalent to convergence" has not been obtained.

The major problem in obtaining this theorem is that of formulating an adequate definition of consistency. In Gear's book there are two definitions for rth order consistency, one for the case $r \leq k$ and another for the case $r > k$. For $r > k$ the definition of rth order consistency is difficult to apply because it is not constructive. And for $r \leq k$ the definition of rth order consistency is not satisfied by all methods having convergence of order r. As an example consider the following multivalue method for first order equations:

$$y_n = y_{n-1} + \frac{h}{2} y_{n-1}' + \frac{h^2}{4} y_{n-1}'' \quad ,$$

$$hy_n' = hf(t_n, y_n) \quad ,$$

$$\frac{h^2}{2} y''_n = hf(t_n,y_n) \quad .$$

The zeroth component of the local truncation error $\underline{\tau}_n$ is defined by

$$y(t_n) + (\underline{\tau}_n)_0 = y(t_{n-1}) + \frac{h}{2} y'(t_{n-1}) + \frac{h^2}{4} y''(t_{n-1}),$$

and since $(\underline{\tau}_n)_0$ is not of order h^2, the method is not consistent in the usual sense. With a change of notation the method can be written

$$y_n = y_{n-1} + \frac{h}{2} y'_{n-1} + \frac{h}{2} \tilde{y}'_{n-1} \quad ,$$

$$hy'_n = hf(t_n,y_n) \quad ,$$

$$h\tilde{y}'_n = hf(t_n,y_n) \quad .$$

This example makes it clear that the method is equivalent to Euler's method with a starting error of order h, and therefore the method is convergent.

In this chapter q-convergence is defined for multivalue methods. The formula used by a multivalue method is written as the sum of a linear part plus a nonlinear part. When $f = f(t)$, the formula becomes a linear difference equation having a unique solution of a special form. The order of the linear part of the formula is defined to be the extent to which the solution

of the difference equation agrees with the correct solution. Also in this chapter it is shown that the numerical solution of a multivalue method is the solution of a corresponding multistep method.

In Chapter 2 the q-root condition and the rth q-order condition are defined, and it is proved for both the corrector only and the $P(EC)^M$ methods that

$$\text{q-convergence like } o(h^{r-1}) \longrightarrow \begin{cases} \text{q-root condition} \\ \text{rth q-order condition} \end{cases} .$$

In Chapter 3 it is shown that

$$\left.\begin{array}{l} \text{q-root condition} \\ \text{rth q-order condition} \end{array}\right\} \longrightarrow \text{q-convergence like } O(h^{r}).$$

Also in Chapter 2 an example is given of a weakly stable (p+2)-value method for pth order equations which is 1-convergent of order 4.

In Chapter 4 the asymptotic form of the error as $h \to 0$ is obtained for a special class of starting values. This result is applied to the case of correct starting values for strongly stable methods. Partial results in this area are obtained by Gear.

Often in practice the stepsize is varied. Let $h_n = t_n - t_{n-1}$. Then the interpolatory technique for varying stepsize is accomplished by premultiplying $\underline{a}_{n-1}$ by the matrix $C_n = \mathrm{diag}(1, h_n/h_{n-1}, \ldots, (h_n/h_{n-1})^{k+p-1})$ before taking the next step. If a method is 1-convergent of order r for fixed stepsize then it is also 1-convergent when the stepsize changes satisfy the following restrictions:

1. $h_n \geq \Delta h$
2. $\sum_{n=1}^{N} |h_n - h_{n-1}| \leq Vh$

where h is the maximum stepsize, Δ and V are fixed positive constants, and N is the number of steps. For $q > 1$, q-convergence of order r is generally not preserved when the stepsize is varied. In fact, almost any method which satisfies the definition of rth order q-convergence for variable stepsize is also 1-convergent of order r. In the second section of Chapter 5, it is shown that a (k+1)-value Adams-Bashforth-Moulton method for first order equations is stable if the stepsize does not change more often than every k-1 steps. The results of Chapter 5 are improvements of work done by Tu (1972).

Section 1.1 The Problem

We are interested in constructing an approximate solution for an initial value problem which conforms to the

Definition of Lipschitz problem:

Let p and q be fixed positive integers where $q \le p$. The pth order q-differential equation $y^{(p)} = f(t,y,\ldots,y^{(p-q)})$ is called a Lipschitz differential equation if

1. f is a continuous function of t on [0,1].
2. f is uniformly Lipschitz continuous in the $y^{(i)}$'s; that is, there exist constants L_0, $L_1,\ldots,L_{p-q}$ such that

$$|f(t,y,\ldots,\tilde{y}^{(i)},\ldots,y^{(p-q)}) - f(t,y,\ldots,y^{(i)},\ldots,y^{(p-q)})| \le L_i|\tilde{y}^{(i)} - y^{(i)}|$$

for all $y,\ldots,\tilde{y}^{(i)},y^{(i)},\ldots,y^{(p-q)}$ and for all $0 \le t \le 1$.

A Lipschitz problem consists of a Lipschitz differential equation together with p initial values $y_0, y_0', \ldots, y_0^{(p-1)}$. □

For notational simplicity, we consider only one differential equation. The theory extends almost immediately to systems of equations if absolute values are

replaced by norms and derivatives like $\partial f/\partial y^{(i)}$ are considered to be Jacobian matrices.

Section 1.2 Polynomial Methods

Polynomial methods are discussed in order to motivate the definition of multivalue methods and the definition of q-convergence.

Let the solution to the Lipschitz problem be $y(t)$. Our aim is to construct approximations $y^i(t;h)$ to $y^{(i)}(t)$ on the interval $[0,1]$ for $i = 0,1,\ldots,p-q$. Higher derivatives of y are not approximated because they are not required for the evaluation of f.

First let us introduce the grid of equidistant points $0 = t_0 < t_1 < \ldots < t_N = 1$ and define the stepsize $h = 1/N$. The approximate solution $y^i(t;h)$ is constructed by piecing together N+1 polynomials each of degree $\leq k+p-1$ where $k \geq 1$:

$$y^i(t;h) = p_n^{(i)}(t) \quad \text{for} \quad t_n - \frac{h}{2} < t \leq t_n + \frac{h}{2} .$$

Three variants of the polynomial method are defined which are analogues of multistep corrector only, $P(EC)^M$, and $P(EC)^ME$ methods. All three methods successively determine a sequence of polynomials $p_1(t)$, $p_2(t),\ldots,p_N(t)$. But first $p_0(t)$ must be obtained by some starting procedure.

In order to justify our definition of convergence, let us write the differential equation as

$$y^{(p-1)}(t) = \int_0^t f(\tau, y(\tau), \ldots, y^{(p-q)}(\tau))d\tau \ .$$

Then it is not necessary for f to be a continuous function of its first argument. At worst, $y^{(p-1)}(t)$ might only be absolutely continuous in which case even for the best polynomial approximation,

$$\max_{0 \le t \le h/2} |p_0(t) - y(t)| = o(h^{p-1})$$

is the strongest statement that can be made.

Definition of q-convergence:

A polynomial method is q-convergent if for any Lipschitz problem,

$$\lim_{h \to 0} y^i(t;h) = y^{(i)}(t) \text{ uniformly in } t$$

for i = 0,1,...,p-q whenever $p_0(t)$ is a function of h satisfying

$$p_0(t) = y(t) + o(h^{p-1}) \text{ uniformly on } [0,h/2]. \quad \square$$

Definition of rth order q-convergence for polynomial methods:

Let r be a positive integer. A polynomial method is q-convergent of order r if for any Lipschitz problem with solution $y(t) \varepsilon C^{r+p}[0,1]$,

$$y^{i}(t;h) = y^{(i)}(t) + O(h^{r}) \text{ uniformly in } t$$

for i = 0,1,...,p-q whenever $p_0(t)$ is a function of h satisfying

$$p_0(t) = y(t) + O(h^{r+p-1}) \text{ uniformly on } [0,h/2]. \quad \square$$

Throughout this thesis the use of the little-oh or the big-oh notation carries with it the assumption that the limit or the bound is uniform with respect to time (t and n). Nevertheless there may be a dependence on the method and the problem.

The accuracy of the approximate solution is limited by the use of polynomials of degree k+p-1. The initial value $y^{(p-1)}(t)$ is approximated on [0,h/2] by a polynomial of degree k, and the error in such an approximation is normally of order h^{k+1}. Therefore no polynomial method has order of q-convergence greater than k+1.

The corrector only variant of the polynomial method defines successive polynomials by

$$p_n(t) = p_{n-1}(t) + \omega_n \, L(\frac{t-t_n}{h})$$

where

$$L(s) \equiv \ell_{k+p-1} \, s^{k+p-1} + \ldots + \ell_1 s + \ell_0$$

and ω_n is chosen so that $p_n(t)$ satisfies the differential equation at $t = t_n$.

Consider the equation $y^{(p)} = f(t)$, and suppose that $p_{n-1}(t)$ is known. To determine $p_n(t)$, ω_n is chosen so that the differential equation is satisfied at $t = t_n$:

$$f(t_n) = p_n^{(p)}(t_n) = p_{n-1}^{(p)}(t_n) + \omega_n \frac{p!}{h^p} \ell_p \; .$$

In general, $p_{n-1}^{(p)}(t_n) \neq f(t_n)$, and so ℓ_p must not be zero. Since multiplying $L(s)$ by a nonzero constant does not change the method, we assume that $L(s)$ is normalized so that $\ell_p = 1$.

The polynomial $L(s)$ determines what information is retained when one step is taken. To be more specific, the ith derivative of $p_{n-1}(t)$ is equal to the ith derivative of $p_n(t)$ at the roots of $L^{(i)}((t-t_n)/h)$.

As an example, consider the explicit k-step Adams-Bashforth method, which uses hy'_{n-1}, $hy'_{n-2}, \ldots,$ hy'_{n-k}, and y_{n-1} to determine y_n. The saved values define an interpolating polynomial $p_{n-1}(t)$ so that $p_{n-1}(t_{n-1}) = y_{n-1}$ and $p'_{n-1}(t_{n-i}) = y'_{n-i}$ for $i = 1,2,\ldots,k$. For explicit methods, y_n is obtained by extrapolation; that is, $y_n = p_{n-1}(t_n)$. Hence $p_n(t_n) = p_{n-1}(t_n)$, and so $L(0) = 0$. Also, since $p_n(t)$ is determined by hy'_n, $hy'_{n-1},\ldots,hy'_{n-k+1}$, and y_n, it follows that $p'_n(t) = p'_{n-1}(t)$ for $t = t_{n-1}, t_{n-2},\ldots,t_{n-k+1}$. Thus $L'(s) = 0$ for $s = -1, -2,\ldots,-(k-1)$. These k conditions are sufficient to determine a normalized polynomial $L(s)$ of degree k.

As a second example, the implicit k-step Adams-Moulton method uses hy'_{n-1}, $hy'_{n-2},\ldots,hy'_{n-k}$, and y_{n-1} to determine y_n. Just as in the first example, $L'(s) = 0$ for $s = -1, -2,\ldots,-(k-1)$. Only one more condition is needed to determine $L(s)$, and that is to choose $L(0)$ so that the formula is of order k+1.

One way of representing the polynomial $p_n(t)$ has been illustrated. Another more convenient representation for $p_n(t)$ is the vector of scaled derivatives

$$\underline{a}_n \equiv \mathrm{col}(p_n(t_n), hp'_n(t_n), \ldots, \frac{h^{k+p-1}}{(k+p-1)!}\, p_n^{(k+p-1)}(t_n)) \ .$$

The method is said to be in normal form if the polynomial is represented by scaled derivatives.

Let us write the method in normal form:

$$(1.1) \qquad \underline{a}_n = \tilde{A}\underline{a}_{n-1} + \omega_n \underline{\ell}$$

where $\tilde{A}$ is the $(k+p)\times(k+p)$ Pascal triangle matrix

$$\begin{pmatrix} 1 & 1 & 1 & \cdot & \cdot & \cdot & 1 \\ & 1 & 2 & \cdot & \cdot & \cdot & k+p-1 \\ & & 1 & \cdot & \cdot & \cdot & \binom{k+p-1}{2} \\ & & & \cdot & & & \cdot \\ & & & & \cdot & & \cdot \\ & 0 & & & & \cdot & \cdot \\ & & & & & & 1 \end{pmatrix}$$

and $\underline{\ell} \equiv \mathrm{col}(\ell_0, \ell_1, \ldots, \ell_{k+p-1})$. (Indexing starts from 0 for all vectors and matrices.) Since ω_n is chosen so that $p_n(t)$ satisfies the differential equation at $t = t_n$,

$$\frac{p!}{h^p} (\underline{a}_n)_p = f(t_n, (\underline{a}_n)_0, \ldots, \frac{(p-q)!}{h^{p-q}} (\underline{a}_n)_{p-q}) \; .$$

Therefore

$$\omega_n = \frac{h^p}{p!} F(t_n, \tilde{A}\,\underline{a}_{n-1} + \omega_n \underline{\ell}) - (\tilde{A}\underline{a}_{n-1})_p$$

where

$$F(t, \underline{a}) \equiv f(t, a_0, \ldots, \frac{(p-q)!}{h^{p-q}} a_{p-q}) \; .$$

A unique solution exists for ω_n if h is sufficiently small so that

$$(1.2) \qquad \sum_{i=0}^{p-q} \frac{i!}{p!} |\ell_i| L_i h^{p-i} < 1 \quad .$$

Note that if $\ell_0 = \ell_1 = \ldots = \ell_{p-q} = 0$ then ω_n is determined explicitly. Thus there are explicit and implicit methods.

For implicit corrector only methods, ω_n is often the solution of a nonlinear equation, which must be solved iteratively. Predictor-corrector polynomial methods are developed for this purpose. For values of h satisfying (1.2), the following iteration converges:

$$\omega_{n,(0)} = 0 \ ,$$

$$\omega_{n,(m+1)} = \frac{h^p}{p!} F(t_n, \tilde{A}\,\underline{a}_{n-1} + \omega_{n,(m)}\underline{\ell}) - (\tilde{A}\underline{a}_{n-1})_p \quad .$$

Define $\underline{a}_{n,(m)} = \tilde{A}\underline{a}_{n-1} + \omega_{n,(m)}\underline{\ell}$, and then the corrector iteration becomes

$$(1.3a) \qquad \underline{a}_{n,(0)} = \tilde{A}\underline{a}_{n-1}$$

$$(1.3b) \qquad \underline{a}_{n,(m+1)} = \underline{a}_{n,(m)} + \underline{\ell}\,[\frac{h^p}{p!} F(t_n, \underline{a}_{n,(m)}) - (\underline{a}_{n,(m)})_p] .$$

In practice only a finite number of corrector iterations

are done, and to simplify the analysis, assume that the number of iterations is a constant M:

$$(1.3c) \qquad \underline{a}_n = \underline{a}_{n,(M)} \, .$$

These last three equations constitute the definition of a $P(EC)^M$ method. The abbreviation stands for "prediction, (evaluation of f, correction)^M."

Note that in general, the polynomial represented by $\underline{a}_{n(M)}$ does not satisfy the differential equation at $t = t_n$. This can be remedied by performing a final correction on the pth component of $\underline{a}_{n,(M)}$:

$$\underline{a}_n = \underline{a}_{n,(M)} + \underline{\delta}_p \left[\frac{h^p}{p!} F(t_n, \underline{a}_{n,(M)}) - (\underline{a}_{n,(M)})_p\right]$$

where $\underline{\delta}_p$ is a column vector with a 1 in the pth position (zero-origin indexing) and zeros elsewhere. This variant of the polynomial method, called a $P(EC)^M E$ method, is not investigated because of the extra complications involved. The analysis is similar to that for $P(EC)^M$ methods.

Section 1.3 Multivalue Methods

Multivalue methods are generalizations of polynomial methods in normal form. A multivalue method is defined either by (1.1) or by (1.3) where $\tilde{A}$ is replaced

by the general matrix A and for a $P(EC)^M$ method, ℓ_p is not required to be 1. Methods with $\ell_p \neq 1$ are less accurate and less stable, and thus they are less likely to have practical value. Both Gear (1971) and Stetter (1973) consider the possibility that $\ell_p \neq 1$, and so for completeness, we do likewise.

The numerical solution produced by a corrector only and a $P(EC)^M$ method satisfy respectively

$$\underline{a}_n = (I - \underline{\ell}\, \underline{\delta}_p^T) A \underline{a}_{n-1} + \underline{\ell}\, \frac{h^p}{p!} F(t_n, \underline{a}_n)$$

and

$$\underline{a}_n = (I - \underline{\ell}\, \underline{\delta}_p^T)^M A \underline{a}_{n-1} + \sum_{m=1}^{M} (I - \underline{\ell}\, \underline{\delta}_p^T)^{M-m}\, \underline{\ell}\, \frac{h^p}{p!} \times$$

$$\times F(t_n, \underline{a}_{n,(m-1)}) \; .$$

Both formulas can be written

$$\underline{a}_n = S \underline{a}_{n-1} + \underline{\tilde{\ell}}\, \frac{h^p}{p!} \Phi(t_n, \underline{a}_{n-1})$$

where

$$S \equiv (I - \underline{\tilde{\ell}}\, \underline{\delta}_p^T) A \; ,$$

$$\underline{\tilde{\ell}} \equiv \begin{cases} \underline{\ell} & \text{for corrector only methods} \\ u_M \underline{\ell} & \text{for } P(EC)^M \text{ methods} \; , \end{cases}$$

and

$$u_m \equiv 1 + (1 - \ell_p) + \ldots + (1 - \ell_p)^{m-1} \; .$$

The dependence of Φ on h is not shown in the notation.

<u>Definition of multivalue method</u>:

Let A be a k+p by k+p matrix and $\underline{\ell}$ be a vector of dimension k+p. A (k+p)-value method uses the formula

$$\underline{a}_n = S\underline{a}_{n-1} + \underline{\tilde{\ell}} \frac{h^p}{p!} \Phi(t_n, \underline{a}_{n-1})$$

where S is defined above and Φ is defined below.

1. Let $0 < \eta < 1$. For a corrector only method, $\ell_p = 1$, $\Phi(t,\underline{a}) = \phi$ where

$$\phi = F(t, S\underline{a} + \underline{\ell} \frac{h^p}{p!} \phi) ,$$

and h satisfies

$$\sum_{i=0}^{p-q} \frac{i!}{p!} |\ell_i| L_i h^{p-i} \le 1 - \eta \quad .$$

2. Let M be a positive integer. For a $P(EC)^M$ method, if $\ell_p = 1$ then $\ell_i \ne 0$ for some $i \le p-q$, and $\Phi(t,\underline{a}) = \phi_{(M)}$ where

$$\phi_{(0)} = h^{-p} p! \, (A\underline{a})_p \quad ,$$

$$\phi_{(m)} = (1 - \frac{u_m}{u_M}) \phi_{(0)} + \frac{1}{u_M} \sum_{j=1}^{m} (1-\ell_p)^{m-j} \quad F(t, \underline{a}_{(j-1)}) \quad ,$$

$$\underline{a}_{(m)} = S\underline{a} + \underline{\tilde{\ell}} \frac{h^p}{p!} \phi_{(m)} \quad . \qquad \square$$

$P(EC)^M$ methods are not permitted to have $\ell_p = 1$ and $\ell_i = 0$ for $i \leq p-q$ because an explicit predictor-corrector method is really a corrector only method if $\ell_p = 1$.

Note that if $0 < \ell_p < 2$ then

$$\underline{\tilde{\ell}} = u_M \underline{\ell} \longrightarrow \ell_p^{-1} \underline{\ell} \qquad \text{as} \quad M \to \infty \qquad .$$

Otherwise u_M diverges as $M \to \infty$, which indicates that an unlimited number of corrector iterations can only be allowed if $0 < \ell_p < 2$.

In the case that $\ell_p = 1$, there is considerable simplifcation in the above definition:

$$\phi_{(0)} = h^{-p} p! (A\underline{a})_p \quad ,$$

$$\phi_{(m)} = F(t, S\underline{a} + \underline{\ell} \frac{h^p}{p!} \phi_{(m-1)}) \quad .$$

It has been noted that using polynomials of degree k+p-1 limits the accuracy of $y(t;h)$. Thus it is reasonable to construct $y(t;h)$ by means of interpolation schemes which use several of the $\underline{a}_n$'s to define $y(t;h)$ at any one point. In order to interpolate, it is necessary to assign a meaning to each component of $\underline{a}_n$. Let the correct value of $\underline{a}_n$ be given by $\underline{a}^c(t_n)$ where $\underline{a}^c(t)$ is defined in terms of the solution $y(t)$ and the stepsize h. If interpolation schemes of

arbitrarily high order are used then the accuracy of the approximate solution is limited only by the accuracy of the computed values $\underline{a}_0$, $\underline{a}_1, \ldots, \underline{a}_N$ (and by the number of points N+1). Convergence should be redefined to require that the computed values $\underline{a}_n$ converge in some sense to the correct values $\underline{a}^c(t_n)$ so that the definition is independent of the techniques used to construct $y(t;h)$. With such a definition, there is no a priori limit on the order of convergence of a method.

In the case of a k-step Adams method, the polynomial $p_n(t)$ is determined by the components of

$$\underline{y}_n \equiv \mathrm{col}(y_n, hy'_n, hy'_{n-1}, \ldots, hy'_{n-k+1}) \ ,$$

and thus there exists some nonsingular matrix T such that $\underline{a}_n = T\underline{y}_n$. Therefore we might define $\underline{a}^c(t) = T\underline{y}(t)$ where

$$\underline{y}(t) = \mathrm{col}(y(t), hy'(t), hy'(t-h), \ldots, hy'(t-[k-1]h)).$$

For k = 2, we would have

$$(1.4) \qquad \underline{a}^c(t) = \begin{pmatrix} y(t) \\ hy'(t) \\ \frac{h}{2}\,(y'(t)-y'(t-h)) \end{pmatrix} .$$

A second possibility is to define $\underline{a}^c(t)$ as the vector of scaled derivatives of $y(t)$. For p = 1 and

and k = 2 we would have

$$\underline{a}^{c}(t) = \begin{bmatrix} y(t) \\ hy'(t) \\ \frac{h^2}{2}\, y''(t) \end{bmatrix} .$$

This differs from (1.4) by a term of order h^3.

For both of the above examples we can write $\underline{a}^{c}(t) = \underline{\Lambda} y(t)$ where $\underline{\Lambda} = \mathrm{col}(\Lambda_0, \Lambda_1, \ldots, \Lambda_{k+p-1})$ is a vector of linear operators. If y(t) is analytic then each Λ_i can be expressed as a power series in the scaled differential operator hD. For example the translation operator can be written as e^{hD} because $e^{hD}y(t) = y(t+h)$. Since $\underline{\Lambda}$ indicates how to interpret the $\underline{a}_n$'s, we define an interpretation to be a vector of linear operators, each component a power series in hD.

From this point on, let us adopt the Nordsieck interpretation

$$\underline{\Lambda} y(t) \equiv \mathrm{col}\left(y(t), hy'(t), \ldots, \frac{h^{k+p-1}}{(k+p-1)!}\, y^{(k+p-1)}(t)\right),$$

and define $\underline{a}(t)$ by

$$a_i(t) = \begin{cases} h^i y^{(i)}(t)/i! & \text{for } i \le r+p-1 \\ 0 & \text{otherwise} \end{cases} .$$

Note that $\underline{a}(t)$ depends implicitly on r, which can be adjusted in case some higher derivatives do not exist.

Before defining q-convergence, let us select some norm $||\cdot||$ for column vectors of dimension k+p. Corresponding norms for matrices and row vectors can be defined in terms of $||\cdot||$. Define the norm $||\cdot||_H = ||H^{-1}\cdot||$ where

$$H \equiv \begin{bmatrix} 1 & & & & & & \\ & h & & & & 0 & \\ & & \cdot & & & & \\ & & & \cdot & & & \\ & & & & h^{p-q} & & \\ & & & & & \cdot & \\ & 0 & & & & & \cdot \\ & & & & & & h^{p-q} \end{bmatrix} .$$

Definition of rth order q-convergence for multivalue methods:

Let r be a positive integer. A multivalue method is q-convergent of order r if for any Lipschitz problem with solution $y(t)\varepsilon C^{r+p}[0,1]$,

$$||\underline{a}_n - \underline{a}(t_n)||_H = O(h^r)$$

whenever $\underline{a}_0$ is a function of h satisfying

$$\underline{a}_0 = \underline{a}(0) + \underline{O}(h^{r+p-1}) \quad .$$

□

It is not difficult to show that this definition is equivalent to the corresponding definition for polynomial methods when $r \leq k+1$.

Section 1.4 Determining the Order of a Method

In the previous section the concept of interpretation was introduced in order to show that there is nothing special about the Nordsieck interpretation. It is clear that the order of convergence depends on the choice of $\underline{\Lambda}$, and this suggests that for any given method we could try to find an interpretation $\underline{\Lambda}^*$ that would make the order of convergence as high as possible. The search for such an interpretation uncovers a constructive definition for the order of a multivalue method. Because $y^{(p-q)}$ appears in the differential equation, it can be shown that the highest order attainable is $k+q+1$. Therefore let us consider only equations of the form $y^{(p)} = f(t)$. The following theorem asserts that for any multivalue method satisfying certain mild conditions, there exists a unique interpretation $\underline{\Lambda}^*$ which makes the method exact whenever $f(t)$ is well-behaved and h is small enough.

The numerical solution of $y^{(p)} = f(t)$ satisfies

$$\underline{a}_n = S\underline{a}_{n-1} + \underline{\tilde{\ell}} \frac{h^p}{p!} f(t_n) \quad . \tag{1.5}$$

Given the solution $y(t)$, the problem is to find $\underline{a}^*(t) = \underline{\Lambda}^* y(t)$ so that

$$\underline{a}^*(t) = S\underline{a}^*(t-h) + \underline{\tilde{\ell}}\,\frac{h^p}{p!}\,y^{(p)}(t) \quad .$$

Write this as the operator equation

$$(1.6) \qquad \underline{\Lambda}^* = Se^{-hD}\underline{\Lambda}^* + \underline{\tilde{\ell}}\,\frac{(hD)^p}{p!} \quad .$$

<u>Theorem 1.1</u>:

Consider a method for which S has no generalized eigenvectors of rank >p corresponding to 1. Then (1.6) has a unique solution

$$\underline{\Lambda}^* = \sum_{j=0}^{\infty} \underline{d}_j \frac{(hD)^j}{j!} \quad ,$$

called the optimal interpretation. For $y(t) \varepsilon C^{\infty}[0,1]$, the power series defining $\underline{a}^*(t) = \underline{\Lambda}^* y(t)$ converges when $h < h_t$ where

$$h_t \equiv \rho\left(\limsup_{j\to\infty} |y^{(j)}(t)|^{1/j}\right)^{-1}$$

and

$$\rho \equiv \min\{|\log\xi| : \xi \neq 1 \text{ is an eigenvalue of } S\} \quad .$$

<u>Proof</u>: Substitute the complex variable z for the operator hD in (1.6) to get

$$(1.7) \qquad S\underline{\lambda}(z) = e^z\underline{\lambda}(z) - \frac{z^p}{p!}\,e^z\underline{\tilde{\ell}}$$

where $\underline{\lambda}(z)$ has been substituted for $\underline{\Lambda}^*$. Let

$$Q_m(z) = q_0 z^m + q_1 z^{m-1} + \dots + q_m$$

be a minimal polynomial for S and then define

$$Q_j(z) = q_0 z^j + q_1 z^{j-1} + \dots + q_j \quad .$$

Repeated applications of (1.7) yield

$$S^j \underline{\lambda}(z) = e^{jz} \underline{\lambda}(z) - \frac{z^p}{p!} \sum_{\ell=1}^{j} e^{\ell z} S^{j-\ell} \underline{\tilde{\ell}}$$

whence

$$Q_m(S) \underline{\lambda}(z) = Q_m(e^z) \underline{\lambda}(z) - \frac{z^p}{p!} \sum_{j=1}^{m} e^{jz} Q_{m-j}(S) \underline{\tilde{\ell}} \quad .$$

The left-hand side vanishes leaving

$$\underline{\lambda}(z) = \frac{z^p}{p! Q_m(e^z)} \sum_{j=1}^{m} e^{jz} Q_{m-j}(S) \underline{\tilde{\ell}} \quad .$$

Because S has no generalized eigenvectors of rank $>p$ corresponding to 1, $Q_m(e^z)$ does not have a zero at $z = 0$ of order $>p$, and so $\underline{\lambda}(z)$ is analytic at $z = 0$. The singularities of $\underline{\lambda}(z)$ are in the set

$$\{\log \xi : \xi \neq 1 \text{ is an eigenvalue of } S\} \quad .$$

Therefore $\underline{\lambda}(z)$ has a MacLaurin series with radius of convergence ρ. □

Consider for example Euler's method:

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \tilde{\underline{\ell}} = \underline{\ell} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad S = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} .$$

The optimal interpretation is

$$\underline{\Lambda}^* = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix} hD + \begin{pmatrix} \frac{1}{6} \\ 0 \end{pmatrix} \frac{(hD)^2}{2} + \begin{pmatrix} -\frac{1}{30} \\ 0 \end{pmatrix} \frac{(hD)^4}{24} + \cdots .$$

$\underline{\Lambda}^* e^t$ is always easy to find, and for this example

$$\underline{\Lambda}^* e^t = \begin{pmatrix} (1 - e^{-h})^{-1} h \\ h \end{pmatrix} e^t .$$

In Section 2.2 the order of a method is defined to be the extent to which the optimal interpretation agrees with the Nordsieck interpretation. For $q > 1$ it is necessary to impose a second condition to ensure that the difference between the numerical solutions with starting values $\underline{a}(0)$ and $\underline{a}^*(0)$ respectively does not grow as n increases. Also for $P(EC)^M$ methods, it is necessary to require that the predictor matrix A be sufficiently close to the Pascal triangle matrix $\tilde{A}$.

Section 1.5 The Relationship to Multistep Methods

Let us modify the proof of Theorem 1.1 in order to show that the numerical solution of a corrector only method satisfies a difference equation of the type that is used by multistep methods. One step of a corrector only method is defined by

$$\underline{a}_n = S\underline{a}_{n-1} + \underline{\ell}\,\frac{h^p}{p!}\,f_n$$

where $f_n = f(t_n, y_n, y_n', \ldots, y_n^{(p-q)})$. Thus

$$S\underline{a}_\nu = \underline{a}_{\nu+1} - \underline{\ell}\,\frac{h^p}{p!}\,f_{\nu+1}$$

and

$$S^j\underline{a}_{n-m} = \underline{a}_{n-m+j} - \sum_{\ell=1}^{j} S^{j-\ell}\,\underline{\ell}\,\frac{h^p}{p!}\,f_{n-m+\ell} .$$

Let $Q_0(z), Q_1(z), \ldots, Q_m(z)$ be the polynomials defined in the proof of Theorem 1.1. Then

$$0 = Q_m(S)\underline{a}_{n-m} = \sum_{j=0}^{m} q_j\underline{a}_{n-j} - \sum_{j=0}^{m-1} Q_j(S)\,\underline{\ell}\,\frac{h^p}{p!}\,f_{n-j} .$$

Since the pth row of S is all zero, S must be singular, and so $q_m = 0$. Therefore

$$(1.8) \qquad \sum_{j=0}^{m-1} q_j\,\frac{h^i}{i!}\,y_{n-j}^{(i)} = \sum_{j=0}^{m-1} (Q_j(S)\,\underline{\ell})_i\,\frac{h^p}{p!}\,f_{n-j}$$

for $i = 0, 1, \ldots, p-q$. What has been derived in (1.8) is the formula for an $(m-1)$-step method where $m \le k+p$.

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

Any polynomial $Q(z) \not\equiv 0$ for which $Q(S) = 0$ could be used instead of $Q_m(z)$, and the numerical solution would satisfy the associated difference equation. Of all difference equations derived in this manner, (1.8) has the lowest order.

Let T be any $(k+p)\times(k+p)$ nonsingular matrix with $\underline{\delta}_i^T T = \underline{\delta}_i^T$ for $i = 0,1,\ldots,p-q,p$, and consider the multivalue method determined by $A' = TAT^{-1}$ and $\underline{\ell}' = T\underline{\ell}$. Then the associated multistep method given by (1.8) remains unchanged even though the order of consistency of the two multivalue methods may differ.

Gear (1971:151) states that "the important characterization of a method is the number of values that are saved from step to step and not the number of steps." Perhaps this should be amended to say that the important characterization of a method is the number of "independent" values that are saved. We have shown that a $(k+p)$-value formula in normal form is equivalent to a $2(k+p-1)$-value multistep formula. In both cases there are $k+p-1$ "independent" values. The "dependent" values are related to these through the differential equation. The importance of the number of "independent" variables is introduced by Gear (1971:141) in connection with stability.

CHAPTER 2

NECESSITY OF THE CONDITIONS FOR q-CONVERGENCE

A method is said to be q-convergent like $o(h^{r-1})$ if it satisfies the definition of rth order q-convergence with $O(h^r)$ replaced by $o(h^{r-1})$; that is, $||\underline{a}_n-\underline{a}(t_n)||_H = o(h^{r-1})$ whenever $\underline{a}_0 = \underline{a}(0) + \underline{O}(h^{r+p-1})$. In the first three sections of this chapter it is shown that a method which is q-convergent like $o(h^{r-1})$ must necessarily satisfy

1. conditions on the eigenvalues of the matrix S.
2. conditions on the accuracy of the method when $f = f(t)$.
3. additional conditions on the accuracy of the method when $f = f(t,y,\ldots,y^{(p-q)})$.

In the fourth section these conditions are applied to study multivalue methods of maximum order.

In Chapter 3 these necessary conditions are shown to be sufficient for q-convergence of order r, and in Chapter 5 this result is extended to some variable step-size methods. For Chapters 2, 3, and 5 the following notation is very useful:

E is the $(k+p)\times(r+p)$ matrix defined by $E_{ij} = \delta_{ij}$,

$\underline{\delta}_j$ is the jth column of E,

$\underline{\varepsilon}_i^T$ is the ith row of E,

P is the $(r+p)\times(r+p)$ Pascal triangle matrix defined by

$$P_{ij} = \binom{j}{i} .$$

Notice that $\underline{a}(t) = E\underline{Y}(t)$ where

$$\underline{Y}(t) \equiv \text{col}(y(t), hy'(t), \ldots, h^{r+p-1}y^{(r+p-1)}(t)/(r+p-1)!) .$$

Also, if $y(t)\varepsilon C^{r+p}[0,1]$ then by employing a truncated Taylor's series about t-h to approximate each component of $\underline{Y}(t)$, it follows that

$$\underline{Y}(t) = P\underline{Y}(t-h) + \underline{O}(h^{r+p}) \quad .$$

Section 2.1 The q-root Condition

The stability properties of a multivalue method depend almost entirely on the matrix S. After all the necessary conditions for q-convergence are proved, it is shown that

$$S = \begin{pmatrix} S_0 & S_1 \\ 0 & S_2 \end{pmatrix}$$

where S_0 is a $(p-q)\times(p-q)$ upper triangular matrix with 1's on the diagonal and S_2 is a $(k+q)\times(k+q)$ matrix whose elementary divisors have the form $(z-\xi)^d$ where either $|\xi| < 1$ or $|\xi| = 1$ with $d \le q$. In this section we prove that if the method is q-convergent then S satisfies the q-root condition as given by the

Definition of q-root Condition:

A multivalue method satisfies the q-root condition if the elementary divisors of S have the form $(z-\xi)^d$ where $|\xi| \le 1$ and if $|\xi| = 1$ then $d \le p$. Furthermore, if there is an elementary divisor $(z-\xi_1)^p$ where $|\xi_1| = 1$ then all other elementary divisors $(z-\xi)^d$ with $|\xi| = 1$ satisfy $d \le q$. ☐

The q-root condition, discovered by Gear, is very similar to the root condition of multistep theory. In fact, the q-root condition implies that the associated difference equation (1.8) satisfies the root condition, which means that the numerical solution of a convergent corrector only method is also the solution of a stable multistep method.

Theorem 2.1:

The q-root condition is necessary for q-convergence like $o(h^{r-1})$.

Proof: Consider the most trivial Lipschitz problem $y^{(p)} = 0$ with $y_0 = y_0' = \dots = y_0^{(p-1)} = 0$. Let $\underline{v}$ be an arbitrary vector and set $\underline{a}_0 = h^{r+p-1}\underline{v}$. The computed solution $\underline{a}_n = h^{r+p-1}S^n\underline{v}$. Hence q-convergence like $o(h^{r-1})$ implies that

$$(2.1) \qquad ||h^p S^n \underline{v}||_H \longrightarrow 0 \quad \text{as} \quad h \longrightarrow 0 \quad .$$

The q-root condition is proved in three parts.

Part I: The eigenvalues of S are on or inside the unit circle. Assume this is false. Then there exists an eigenvector $\underline{x}$ corresponding to an eigenvalue ξ where $|\xi| > 1$. Hence $S^N\underline{x} = \xi^N\underline{x}$ and so $||h^p S^N \underline{x}||_H \to \infty$ as $h \to 0$, which contradicts (2.1).

Part II: An eigenvalue of S on the unit circle may not correspond to an elementary divisor of degree greater than p. Assume this is false. Then there exists a generalized eigenvector $\underline{x}_p$ of rank p+1 corresponding to an eigenvalue ξ where $|\xi| = 1$. Defining $\underline{x}_i = (S-\xi I)^{p-i}\underline{x}_p$ allows us to write

$$S^N\underline{x}_p = \binom{N}{0} \xi^N \underline{x}_p + \binom{N}{1} \xi^{N-1}\underline{x}_{p-1} + \dots + \binom{N}{p} \xi^{N-p}\underline{x}_0$$

whence

$$h^p S^N \underline{x}_p = \frac{\xi^{N-p}}{p!} \underline{x}_0 + \underline{O}(h) \quad .$$

Therefore $||h^p S^N \underline{x}_p||_H$ does not tend to zero as h → 0, which contradicts (2.1).

Part III: If there is an elementary divisor $(z-\xi_1)^p$ where $|\xi_1| = 1$ then all other elementary divisors $(z-\xi)^d$ with $|\xi| = 1$ satisfy $d \leq q$. Assume this is false. Let $\underline{x}_{p-1}$ be a generalized eigenvector of rank p corresponding to ξ_1, and let $\underline{w}_q$ be a generalized eigenvector of rank q+1 corresponding to an eigenvalue ξ on the unit circle such that $\underline{x}_0, \underline{x}_1, \ldots, \underline{x}_{p-1}, \underline{w}_0, \underline{w}_1, \ldots, \underline{w}_q$ are linearly independent. Here $\underline{x}_i \equiv (S-\xi_1 I)^{p-1-i} \underline{x}_{p-1}$ and $\underline{w}_i \equiv (S-\xi I)^{q-i} \underline{w}_q$. For an arbitrary vector $\underline{v}$, (2.1) implies that

$$(2.2) \qquad (S^N \underline{v})_i = o(N^q) \qquad \text{if} \quad i \geq p-q \quad .$$

Clearly

$$S^N \underline{x}_{p-1} = \binom{N}{p-1} \xi_1^{N-p-1} \underline{x}_0 + \ldots + \binom{N}{q} \xi_1^{N-q} \underline{x}_{p-q} + \underline{o}(N^q) \quad .$$

If this is substituted into (2.2) with $\underline{v} = \underline{x}_{p-1}$ then it is not difficult to see that

$$(\underline{x}_0)_i = (\underline{x}_1)_i = \ldots = (\underline{x}_{p-q-1})_i = 0 \quad \text{for} \quad i \geq p-q \ .$$

Similarly

$$S^N \underline{w}_q = \binom{N}{q} \xi^{N-q} \underline{w}_0 + o(N^q) \quad ,$$

and after substituting this into (2.2) with $\underline{v} = \underline{w}_q$, we conclude that

$$(\underline{w}_0)_i = 0 \qquad \text{for} \quad i \geq p-q \; .$$

Therefore each of $\underline{x}_0, \ldots, \underline{x}_{p-q-1}, \underline{w}_0$ is a linear combination of $\underline{\delta}_0, \underline{\delta}_1, \ldots, \underline{\delta}_{p-q-1}$, which is impossible because $\underline{x}_0, \ldots, \underline{x}_{p-q-1}, \underline{w}_0$ are linearly independent. □

Section 2.2 The Accuracy of the Corrector

In this section two conditions are derived which are necessary for convergence when $f = f(t)$: the corrector condition and the growth condition. These two conditions determine the order of the linear part of the formula, $S\underline{a}_{n-1}$.

In the next section the predictor condition is derived, which is necessary for convergence when $f = f(t, y, \ldots, y^{(p-q)})$. The predictor condition determines the order of the nonlinear part of the formula, $\underline{\tilde{\ell}} \frac{h^p}{p!} \Phi(t_n, \underline{a}_{n-1})$.

These conditions are combined in the

Definition of rth q-order condition:

Let r be a positive integer. A multivalue method satisfies the rth q-order condition if S has no generalized eigenvector of rank >p corresponding to 1 and if the method satisfies

1. the corrector condition:

$$D_{ij} = \delta_{ij} \qquad \text{for } j < \min(i,p-q)+r$$

where D is the $(k+p)\times(r+p)$ matrix defined by $D_{ij} = (\underline{d}_j)_i$ or equivalently $D = \text{row}\,(\underline{d}_0, \underline{d}_1, \ldots, \underline{d}_{r+p-1})$ and the $\underline{d}_j$'s are defined by Theorem 1.1.

2. the growth condition: for $r+p-q \leq j < r+p-1$ and $i \geq p-q$,

$$(S^n(\underline{d}_j - \underline{\delta}_j))_i = O(n^{j-r-p+q}) \quad .$$

3. the predictor condition: for a $P(EC)^M$ method if $\ell_p = 1$ then

$$A_{pj} = \binom{j}{p} \qquad \text{for } j < r+p-Mc$$

where c is the smallest integer $\geq q$ such that $\ell_{p-c} \neq 0$. □

It is easy to see that the (r+1)th condition implies the rth q-order condition. The corrector condition states

that D has the form

$$\begin{array}{ccccccccccc|c}
 & & & \text{col } r & & & & & \text{col } r+p-q & & \text{col } r+p-1 & \\
1 & & & x & \cdots & \cdots & \cdots & \cdots & x & \cdots & x & \\
 & & & & \ddots & & & & \vdots & & \vdots & \\
 & 1 & & & & \ddots & & & \vdots & & \vdots & \\
 & & \ddots & & & & \ddots & & \vdots & & \vdots & \\
 & & \ddots & & 0 & & & x & x & \cdots & x & \\
 & & \ddots & & & & & & & & & \\
 & & & \ddots & & & & & x & \cdots & x & \text{row } p-q \\
 & & & & \ddots & & & & \vdots & & \vdots & \\
 & & & & & \ddots & & & \vdots & & \vdots & \\
 & & & & & & \ddots & & \vdots & & \vdots & \\
 & 0 & & & & & & 1 & \vdots & & \vdots & \\
 & & & & & & & & \vdots & & \vdots & \\
 & & & & & & & & x & \cdots & x &
\end{array}$$

This implies that $||\underline{a}^*(t)-\underline{a}(t)||_H = O(h^r)$ for small h. For $q = 1$ the growth condition does not apply; otherwise it ensures that $||S^n(\underline{a}(0)-\underline{a}^*(0))||_H = O(h^r)$.

The necessity of parts 1 and 2 of the rth q-order condition is proved by considering the problem $y^{(p)} = e^{\lambda t}$ with $y_0^{(i)} = \lambda^i$ where $|\lambda| \leq 1$. Then $\underline{a}^*(t)$ exists for $h < \rho$. Since $\underline{a}^*(t_0), \underline{a}^*(t_1), \ldots, \underline{a}^*(t_N)$ is a numerical solution,

it follows that for any other numerical solution $\underline{a}_0, \underline{a}_1, \ldots, \underline{a}_N$,

$$\underline{a}_n - \underline{a}^*(t_n) = S^n(\underline{a}_0 - \underline{a}^*(0)) \quad . \tag{2.3}$$

<u>Theorem 2.2</u>:

Parts 1 and 2 of the rth q-order condition are necessary for q-convergence like $o(h^{r-1})$.

<u>Proof</u>: By definition,

$$\underline{a}^*(t) = \sum_{j=0}^{\infty} \underline{d}_j \frac{h^j y^{(j)}(t)}{j!}$$

$$= D\underline{Y}(t) + \underline{o}(h^{r+p-1}) \quad . \tag{2.4}$$

Let the starting value be correct: $\underline{a}_0 = \underline{a}(0) = E\underline{Y}(0)$. q-convergence like $o(h^{r-1})$ means that

$$\underline{a}_N = \underline{a}(1) + H\underline{o}(h^{r-1})$$

$$= E\underline{Y}(1) + H\underline{o}(h^{r-1}) \quad . \tag{2.5}$$

Substituting (2.4) and (2.5) into (2.3) with n = N gives

$$E\underline{Y}(1) + H\underline{o}(h^{r-1}) - D\underline{Y}(1) - \underline{o}(h^{r+p-1})$$

$$= S^N(E\underline{Y}(0) - D\underline{Y}(0) - \underline{o}(h^{r+p-1})) \quad .$$

It follows from (2.1) in the proof of Theorem 2.1 that $h^p S^N \underline{v} = H\underline{o}(1)$ for any arbitrary vector $\underline{v}$, and hence $S^N \underline{o}(h^{r+p-1}) = H\underline{o}(h^{r-1})$. Therefore

$$S^N (D-E)\underline{Y}(0) - (D-E)\underline{Y}(1) = H\underline{o}(h^{r-1}) \ ,$$

or equivalently

$$\sum_{j=0}^{r+p-1} \lambda^j S^N \frac{h^j}{j!} (\underline{d}_j - \underline{\delta}_j) - \sum_{j=0}^{r+p-1} \lambda^j e^\lambda \frac{h^j}{j!} (\underline{d}_j - \underline{\delta}_j) = H\underline{o}(h^{r-1}) \ .$$

The functions $1, \lambda, \ldots, \lambda^{r+p-1}$, e^λ, $\lambda e^\lambda, \ldots, \lambda^{r+p-1} e^\lambda$ are linearly independent on $[-1,1]$, and so there exists a set J of $2(r+p)$ points in $[-1,1]$ such that these functions are linearly independent on J. Therefore each of the $2(r+p)$ coefficients of these functions must equal $H\underline{o}(h^{r-1})$. Thus

$$D_{ij} = \delta_{ij} + o(h^{r-1-j+\min(i,p-q)})$$

from which the corrector condition follows. Also for $i \geq p-q$

$$(S^N(\underline{d}_j - \underline{\delta}_j))_i = o(N^{j-r+1-p+q})$$

where N^{-1} has been substituted for h. Since the eigenvalues of S are on or inside the unit circle, either the left-hand side converges to zero as $N \to \infty$ or there is some

nonnegative integer m such that $N^{-m}(S^N(\underline{d}_j-\underline{\delta}_j))_i$ is a bounded nonconvergent sequence. From this it follows that the right-hand side can be replaced by $O(N^{j-r-p+q})$ if $j \geq r+p-q$ thus proving the growth condition. □

Section 2.3 The Accuracy of the Predictor

This section applies only to $P(EC)^M$ methods. Recall that the definition of a $P(EC)^M$ method with $\ell_p=1$ required that $\ell_i \neq 0$ for some $i \leq p-q$. This restriction was made to ensure the existence of c, which is the smallest integer $\geq$ q such that $\ell_{p-c} \neq 0$. If $c>q$ then the method is partially explicit because the predicted value of $y^{(p-q)}$ is left unchanged by subsequent corrections.

The predictor condition requires that $y_{n,(0)}$, $y'_{n,(0)},\ldots,y^{(p-q)}_{n,(0)}$ have order of accuracy $r-Mc$, and it suggests that each correction increases the order by c. If $\ell_p \neq 1$, the predicted values have order r-q and this is not increased until the last correction, when it is increased by q. Gear (1971:191) states in one of his proofs that if the corrector has order r for 1-differential equations then the predictor must have order r-M. This statement is not true if $\ell_p = 1$ and $\ell_{p-1} = 0$.

To prove the necessity of the predictor condition, we consider the problem $y^{(p)} = y^{(p-c)}$ with $y_0 = \ldots = y_0^{(p-1)} = 1$.

Theorem 2.3:

Part 3 of the rth q-order condition is necessary for q-convergence like $o(h^{r-1})$.

Proof: Assume $\ell_p = 1$ and $r+p-Mc > 0$; otherwise there is nothing to prove. The major thrust of the proof consists of arriving at a contradiction by assuming that the predictor condition is violated. That is, a contradiction is obtained by assuming that an integer $s < r$ exists such that

$$A_{pj} = P_{pj} \qquad \text{for} \quad j < s+p-Mc \tag{2.6}$$

$$A_{p,s+p-Mc} = P_{p,s+p-Mc} + \alpha \tag{2.7}$$

where $\alpha \neq 0$. Let us consider the following two initial value problems:

1. $y^{(p)} = e^t$ with $y_0 = \ldots = y_0^{(p-1)} = 1$

2. $y^{(p)} = y^{(p-c)}$ with $y_0 = \ldots = y_0^{(p-1)} = 1$,

which have the same solution $y(t) = e^t$. By the hypothesis the numerical solution of problem 1 satisfies

$$\underline{a}_n' = E\underline{Y}(t_n) + H\underline{o}(h^{r-1})$$

and the numerical solution of problem 2 satisfies

$$(2.8) \qquad \underline{a}_n = E\underline{Y}(t_n) + H\underline{o}(h^{r-1}) \quad .$$

Subtracting these equations shows that the difference $\underline{c}_n \equiv \underline{a}_n - \underline{a}'_n$ satisfies

$$\underline{c}_n = H\underline{o}(h^{r-1}) \quad ,$$

and in particular

$$(\underline{c}_N)_0 = o(h^{r-1}) \quad .$$

The remaining part of the proof consists of finding recurrence relations that are satisfied by the numerical solutions $\underline{a}_n$ and $\underline{a}'_n$. Then by using assumptions (2.6) and (2.7), it is shown that the difference $(\underline{c}_N)_0$ satisfies

$$(2.9) \qquad (\underline{c}_N)_0 = h^s \beta (e - \sum_{i=0}^{p-1} \frac{1}{i!}) + o(h^s)$$

where

$$\beta = \frac{\alpha p!}{(s+p-Mc)!} \left(\frac{(p-c)!}{p!} \ell_{p-c}\right)^M$$

with $\alpha \neq 0$ and $\ell_{p-c} \neq 0$. The two equations for $(\underline{c}_N)_0$ are contradictory because

$$\lim_{h\to 0} h^s \beta (e - \sum_{i=0}^{p-1} \frac{1}{i!})/h^{r-1} \neq 0 \quad .$$

Thus assumptions (2.6) and (2.7) are false and the method satisfies the predictor condition. To complete the details of the proof it must be shown that

$$(2.10)\qquad \underline{a}_n' = S\underline{a}_n' + \underline{\ell}\,\frac{h^p}{p!}\,e^{nh}$$

$$(2.11)\qquad \underline{a}_n = S\underline{a}_n + \underline{\ell}\,\frac{h^p}{p!}\,(e^{nh} + h^s\beta e^{nh} + o(h^s))$$

and thus

$$\underline{c}_n = S\underline{c}_{n-1} + \underline{\ell}\,\frac{h^p}{p!}\,(h^s\beta e^{nh} + o(h^s)) \quad .$$

This recurrence for $\underline{c}_n$ yields

$$\underline{c}_n = h^s \sum_{j=1}^{n} S^{n-j}\,\underline{\ell}\,\frac{h^p}{p!}\,\beta e^{jh} + \sum_{j=1}^{n} o(h^{s+p})\,S^{n-j}\underline{\ell} \quad .$$

Once equations (2.10) and (2.11) have been verified, it is easily checked that the first summation is the nth term of the numerical solution of the problem

$$\delta^{(p)}(t) = \beta e^t \quad \text{with} \quad \delta_0 = \ldots = \delta_0^{(p-1)} = 0 \ .$$

Since the method is q-convergent

$$(\underline{c}_n)_0 = h^s(\delta(t_n) + o(1)) + \sum_{j=1}^{n} o(h^{s+p})\,(S^{n-j}\underline{\ell})_0 \quad .$$

Also, by Theorem 2.1 the method satisfies the q-root condition and thus $(S^{n-j}\underline{\ell})_0 = O(h^{1-p})$. Therefore

$$(\underline{c}_n)_0 = h^s \delta(t_n) + o(h^s) \ ,$$

which yields (2.9). Equation (2.10) may be verified by using (1.5). This leaves us to do the <u>Verification of (2.11)</u>: By definition

$$(2.12) \qquad \underline{a}_n = S\underline{a}_{n-1} + \underline{\ell}\,\frac{h^p}{p!}\,\phi_{n,(M)}$$

where

$$(2.13) \qquad \phi_{n,(0)} = h^{-p}\,p!\,\underline{\delta}_p^T A\underline{a}_{n-1} \ ,$$

$$(2.14) \qquad \phi_{n,(m)} = h^{-p+c}(p-c)!\,\underline{\delta}_{p-c}^T(S\underline{a}_{n-1} + \underline{\ell}\,\frac{h^p}{p!}\,\phi_{n,(m-1)}) .$$

Using $\ell_p = 1$, it follows that $(I - \underline{\ell}\,\underline{\delta}_p^T)\underline{\ell} = \underline{0}$, and thus $(I - \underline{\ell}\,\underline{\delta}_p^T)^M = I - \underline{\ell}\,\underline{\delta}_p^T$. Therefore multiplying (2.12) by $I - \underline{\ell}\,\underline{\delta}_p^T$ gives

$$\begin{aligned} S\underline{a}_{n-1} &= (I - \underline{\ell}\,\underline{\delta}_p^T)\underline{a}_n \\ &= E\underline{Y}(t_n) - \underline{\ell}\,\frac{h^p}{p!}\,e^{nh} + H\underline{o}(h^{r-1}) \quad \text{by} \quad (2.8). \end{aligned}$$

Substitute this into (2.14) to get

$$(2.15) \qquad \phi_{n,(m)} - e^{nh} = h^c\,\frac{(p-c)!}{p!}\,\ell_{p-c}(\phi_{n,(m-1)} - e^{nh}) + o(h^{r-1}) .$$

From the initial assumptions (2.6) and (2.7)

$$(\underline{\delta}_p^T AE)_j = ((\underline{\varepsilon}_p^T + \alpha\underline{\varepsilon}_{s+p-Mc}^T)P)_j \qquad \text{for} \quad j \le s+p-Mc$$

[illegible]

[illegible]

$$[illegible]$$

[illegible]

$$[illegible]$$

[illegible]

$$[illegible]$$

[illegible]

$$[illegible]$$

[illegible]

$$[illegible]$$

[illegible]

$$[illegible]$$

whence

$$\underline{\delta}_p^T AE\underline{Y}(t_{n-1}) = \frac{h^p}{p!}\, e^{nh} + \alpha \frac{h^{s+p-Mc}}{(s+p-Mc)!}\, e^{nh} + O(h^{s+p-Mc+1}) \ .$$

Therefore from (2.8) and (2.13)

$$\begin{aligned} \phi_{n,(0)} &= h^{-p} p!\, \underline{\delta}_p^T A(E\underline{Y}(t_{n-1}) + H\underline{o}(h^{r-1})) \\ &= e^{nh} + \alpha \frac{h^{s-Mc}\ \ p!}{(s+p-Mc)!}\, e^{nh} + O(h^{s-Mc+1}) \\ &\qquad + o(h^{-p+\min(p-q,s+p-Mc)+r-1}) \ , \end{aligned}$$

and so

$$\phi_{n,(0)} - e^{nh} = \frac{h^{s-Mc}\ \ \alpha p!}{(s+p-Mc)!}\, e^{nh} + o(h^{s-Mc}) \ . \tag{2.16}$$

Solve for $\phi_{n,(M)} - e^{nh}$ using (2.15) and (2.16):

$$\phi_{n,(M)} - e^{nh} = h^s \beta e^{nh} + o(h^s) \ .$$

The proof of (2.11) follows from this and (2.12). □

Section 2.4 Methods of Maximum Order

The following lemma partially characterizes the S matrix when the rth q-order condition is satisfied. Furthermore it is used to gain some insight into the vectors $\underline{d}_0, \underline{d}_1, \dots, \underline{d}_{p-1}$.

Lemma 2.1:

If S has no generalized eigenvector of rank >p corresponding to 1 then

$$SD = DP - \underline{\tilde{\ell}}\ \underline{\varepsilon}_p^T P \quad . \tag{2.17}$$

Proof: From (1.7) in the proof of Theorem 1.1,

$$S\underline{\lambda}(z) = e^z\underline{\lambda}(z) - \frac{z^p}{p!}\, e^z\underline{\tilde{\ell}}$$

with

$$\underline{\lambda}(z) = \sum_{j=0}^{\infty} \underline{d}_j \frac{z^j}{j!} \quad .$$

Take the jth derivative where $0 \leq j \leq r+p-1$:

$$S\underline{\lambda}^{(j)}(z) = e^z \sum_{i=0}^{j} \binom{j}{i} \underline{\lambda}^{(i)}(z) - \underline{\tilde{\ell}}\, e^z \sum_{i=0}^{j} \binom{j}{i} \frac{d^i}{dz^i} \frac{z^p}{p!} \quad .$$

Set z = 0 to get

$$S\underline{d}_j = \sum_{i=0}^{j} \binom{j}{i} \underline{d}_j - \binom{j}{p} \underline{\tilde{\ell}}$$

thus verifying the jth column of (2.17). □

Corollary:

If a method satisfies the q-root condition and the rth q-order condition then

$$S_{ij} = ((I - \underline{\tilde{\ell}}\ \underline{\delta}_p^T)\tilde{A})_{ij} \quad \text{for} \quad j < \min(i+1, p-q)+r \ .$$

<u>Proof</u>: Rearrange Lemma 2.1 to get

$$SE = (I - \underline{\tilde{\ell}}\ \underline{\delta}_p^T)EP + (D-E)P - S(D-E) \ .$$

Column by column it can be shown that S has the required form. □

If the hypotheses of the corollary are satisfied with r = 1, then S has the form

$$\begin{pmatrix}
1 & 1 & x & \cdots & x & x & \cdots & x \\
 & 1 & 2 & \cdots & x & x & \cdots & x \\
 & & 1 & \ddots & \vdots & \vdots & & \vdots \\
 & & & \ddots & p-q & x & \cdots & x \\
 & & & & 1 & x & \cdots & x \\
 & 0 & & & & \vdots & & \vdots \\
 & & & & & x & \cdots & x
\end{pmatrix}$$

Consider a method which is q-convergent. By Theorem 2.1 the q-root condition must hold, and so S has no generalized eigenvectors of rank >p corresponding to 1. Hence by the last equation in the proof of Lemma 2.1,

$$(S - I)\underline{d}_j = \sum_{i=0}^{j-1} \binom{j}{i} \underline{d}_j \qquad \text{for} \quad j \le p-1 \ ,$$

or in other words

$$(S - I)\underline{d}_j \ \varepsilon \ \langle \underline{d}_0, \underline{d}_1, \ldots, \underline{d}_{j-1} \rangle$$

where angle brackets denote the linear span of the enclosed vectors. From this it follows that for $j = 0,1,\ldots,p-1$, $\underline{d}_j$ is a generalized eigenvector of rank $\le j+1$ corresponding to 1.

The next lemma shows that it is not necessary to determine D in order to verify the corrector condition and the growth condition. Any matrix $\hat{D}$ which satisfies (2.19) can be used in place of D in the rth q-order condition.

<u>Lemma 2.2</u>:

Assume a method satisfies the q-root condition, and suppose that there exists a $(k+p)\times(r+p)$ matrix $\hat{D}$ such that

$$\text{(2.18)} \qquad \hat{D}_{ij} = \delta_{ij} \qquad \text{for } j < \min(i,p-q)+r \ ,$$

$$\text{(2.19)} \qquad S\hat{D} = \hat{D}P - \underline{\tilde{\ell}}\ \underline{\varepsilon}_p^T P \ ,$$

and

$$(S^n(\hat{D}-E))_{ij} = O(n^{j-r-p+q})$$

for $r+p-q \leq j < r+p-1$ and $i \geq p-q$. Then the method satisfies parts 1 and 2 of the rth q-order condition.

Proof: The proof is divided into two parts. The first part establishes two important facts (equations (2.20) and (2.21)) about certain subspaces invariant under S. The second part actually proves the statement of the lemma.

Part I: Let X_{p-1} be the subspace of generalized eigenvectors of S corresponding to 1 and define the subspaces

$$X_i = (S-I)^{p-1-i} X_{p-1} \quad .$$

Let $\underline{\hat{d}}_j = \hat{D}\underline{\varepsilon}_j$. We assert that

$$\underline{\hat{d}}_j \varepsilon X_j \qquad \text{for} \quad j \leq p-1 \quad . \tag{2.20}$$

This assertion is verified by successively proving the following statements:

$\underline{\hat{d}}_0 \varepsilon X_{p-1}, \underline{\hat{d}}_1 \varepsilon X_{p-1}, \underline{\hat{d}}_2 \varepsilon X_{p-1}, \ldots, \underline{\hat{d}}_{p-1} \varepsilon X_{p-1}$;

$\underline{\hat{d}}_0 \varepsilon X_{p-2}, \underline{\hat{d}}_1 \varepsilon X_{p-2}, \ldots, \underline{\hat{d}}_{p-2} \varepsilon X_{p-2}$;

$\vdots$

$\underline{\hat{d}}_0 \varepsilon X_1, \ \underline{\hat{d}}_1 \varepsilon X_1$;

$\underline{\hat{d}}_0 \varepsilon X_0$.

In order to prove that $\hat{\underline{d}}_i \varepsilon X_{p-1}$, use the following information:

1. the ith column of (2.19):

$$(S-I)\hat{\underline{d}}_i = \binom{i}{i-1}\hat{\underline{d}}_{i-1} + \dots + \binom{i}{0}\hat{\underline{d}}_0 \quad .$$

2. $\hat{\underline{d}}_0, \dots, \hat{\underline{d}}_{i-1} \varepsilon X_{p-1}$.

In order to prove that $\hat{\underline{d}}_i \varepsilon X_j$ where $j < p-1$, use the following information:

1. the (i+1)th column of (2.19):

$$(S-I)\hat{\underline{d}}_{i+1} = \binom{i+1}{i}\hat{\underline{d}}_i + \binom{i+1}{i-1}\hat{\underline{d}}_{i-1} + \dots + \binom{i+1}{0}\hat{\underline{d}}_0 \; .$$

2. $\hat{\underline{d}}_{i+1} \varepsilon X_{j+1}$.

3. $\hat{\underline{d}}_0, \dots, \hat{\underline{d}}_{i-1} \varepsilon X_j$.

It is clear from (2.18) that for $j \le p-q$

$$\hat{\underline{d}}_j = \underline{\delta}_j + \text{some vector in } \langle \underline{\delta}_0, \dots, \underline{\delta}_{j-r} \rangle \; .$$

By induction on j it follows that for $j \le p-q$

$$\begin{aligned} \langle \underline{\delta}_0, \underline{\delta}_1, \dots, \underline{\delta}_j \rangle &\subseteq \langle \hat{\underline{d}}_0, \hat{\underline{d}}_1, \dots, \hat{\underline{d}}_j \rangle \\ &\subseteq X_0 \cup X_1 \cup \dots \cup X_j \qquad \text{by (2.20)} \\ &\subseteq X_j \quad . \end{aligned}$$

Since $X_0 \neq \{\underline{0}\}$, there is a generalized eigenvector $\underline{x}_{p-1}$ of rank p corresponding to 1. Let $\underline{x}_i = (S-I)^{p-1-i}\underline{x}_{p-1}$. By the q-root condition, X_{p-1} is the direct sum of $<\underline{x}_q,\underline{x}_{q+1},\ldots,\underline{x}_{p-1}>$ and the subspace of generalized eigenvectors of rank $\leq q$. Thus $\dim X_{p-q-1} = \dim (S-I)^q X_{p-1} \leq p-q$, and since $\dim X_{j-1} < \dim X_j$, we have $\dim X_j \leq j+1$ for $j < p-q$. Therefore

$$(2.21) \qquad X_j = <\underline{\delta}_0,\underline{\delta}_1,\ldots,\underline{\delta}_j> \qquad \text{if } j < p-q\ .$$

<u>Part II</u>: Subtract (2.19) from (2.17)

$$S(D-\hat{D}) = (D-\hat{D})P \quad .$$

Let $\underline{v}_j = (D-\hat{D})\underline{\varepsilon}_{j+r}$ for $j = -r,-r+1,\ldots,0,\ldots,p-1$. Then by the same argument as in (2.20) it is possible to show that

$$\underline{v}_j \varepsilon X_j \ .$$

By (2.21), $\underline{v}_j \ \varepsilon < \underline{\delta}_0,\underline{\delta}_1,\ldots,\underline{\delta}_j >$ for $j < p-q$ so that

$$(\underline{v}_j)_i = 0 \qquad \text{if} \quad i > j \text{ and } j < p-q\ .$$

Equivalently

$$D_{ij} - \hat{D}_{ij} = 0 \quad \text{for } j < \min(i,p-q)+r \quad .$$

The corrector condition follows from this and (2.18). Let $p-q \leq j < p-1$. Since $\underline{v}_j \varepsilon X_j$, there exist vectors $\underline{x}_i \varepsilon X_i$ such that

$$S^n \underline{v}_j = \underline{v}_j + n\underline{x}_{j-1} + \dots + \binom{n}{j-p+q} \underline{x}_{p-q}$$

$$+ \text{ some vector in } <\underline{\delta}_0, \underline{\delta}_1, \dots, \underline{\delta}_{p-q-1}> .$$

Therefore when $i \geq p-q$

$$\begin{aligned} (S^n(D-E))_{ij} &= (S^n(D-\hat{D}))_{ij} + (S^n(\hat{D}-E))_{ij} \\ &= (S^n \underline{v}_{j-r})_i + O(n^{j-r-p+q}) \\ &= O(n^{j-r-p+q}) \quad , \end{aligned}$$

which proves the growth condition. □

As an example it can be shown that parts 1 and 2 of the rth q-order condition are satisfied when

$$A_{ij} = \binom{j}{i} \quad \text{for} \quad j \leq r+p-1 \quad .$$

Simply apply Lemma 2.2 with $\hat{D} = E$. Recall that the derivation of the multivalue method starts with the Pascal triangle matrix $\tilde{A}$.

Gear (1971:203) shows that there exist strongly stable (k+p)-value methods which are q-convergent of order k+q. A strongly stable method has no extraneous roots on the unit circle. In fact if $q = 1$, $\underline{\ell}$ can be chosen so that the extraneous roots are all zero. Gear (1971:154) gives coefficients for some of these higher

order analogues of Adams' method. The case $p = 1$ and $k = 5$ is Nordsieck's method.

There are no $(k+p)$-value methods which are q-convergent of order $k+q+1$ when $k+q+1$ is odd for the cases $p = q = 1$ and $p = q = 2$. Let us prove this statement by assuming that such a method exists. Set $r = k+q+1$ in Lemma 2.1. If the corrector condition is satisfied, the $(p,k+p)$th component of (2.17) forces $\ell_p = 1$. Clearly the order of a $P(EC)^M$ method with $\ell_p = 1$ cannot exceed the order of the corresponding corrector only method, and to every corrector only method there corresponds a stable $(k+q-1)$-step method. However, for the two cases under consideration it is known (Henrici (1962)) that the order of a stable $(k+q-1)$-step method cannot exceed $k+q$ if $k+q-1$ is odd.

There exist $(k+p)$-value methods which are q-convergent of order $k+q+1$ when $k+q+1 = 4$ and $q = 1$, or equivalently, there exist $(p+2)$-value methods which are 1-convergent of order 4. To show this, we use Lemma 2.2 with

$$A = \tilde{A} \qquad \ell_i = \binom{p+2}{p}^{-1} \binom{p+2}{i}$$

$$\hat{D} = \left[\begin{array}{cccccc|c} 1 & & & & & & \\ & & & 0 & & & \\ & 1 & & & & & \underline{\hat{d}} \\ & & \cdot & & & & \\ & & & \cdot & & & \\ 0 & & & & \cdot & & \\ & & & & 1 & 0 & \end{array}\right]$$

Then all but the last column of (2.19) are automatically satisfied. The last column of this equation can be written in the form

$$\left[\begin{array}{cccccc} 0 & x & \cdot\;\cdot\;\cdot & & x & x \\ & & \cdot & & \vdots & \vdots \\ & & & \cdot & \vdots & \vdots \\ & & & & x & x \\ & 0 & & & -1 & 0 \\ & & & & x & x \end{array}\right] \left[\begin{array}{c} \hat{d}_0 \\ \vdots \\ \hat{d}_{p-1} \\ \hat{d}_p \\ \hat{d}_{p+1} \end{array}\right] = \left[\begin{array}{c} x \\ \vdots \\ x \\ 0 \\ x \end{array}\right]$$

Clearly $\hat{d}_p = 0$. It so happens that the same value of $\hat{d}_{p+1}$ satisfies both the (p-1)th and the (p+1)th equation. The other components $\hat{d}_{p-1}, \hat{d}_{p-2}, \ldots, \hat{d}_1$ can be determined by back substitution. The value of $\hat{d}_0$ is immaterial. This method is weakly stable because the extraneous eigenvalues are 0 and -1.

Assume some method satisfies the (k+q+2)th q-order condition. With $r = k+q+2$, the $(p-1, k+p)$th component of (2.17) is

$$\binom{k+p}{p}\tilde{\ell}_{p-1} = \binom{k+p}{p-1}$$

and the (p-1,k+p-1)th component is

$$\binom{k+p+1}{p}\tilde{\ell}_{p-1} = \binom{k+p+1}{p-1} .$$

Because these two equations are contradictory, no method satisfies the (k+q+2)th q-order condition.

CHAPTER 3

SUFFICIENCY OF THE CONDITIONS FOR q-CONVERGENCE

In the first section of this chapter q-stability and rth order q-consistency are defined, and it is shown that together these properties are sufficient to prove that a method is rth order q-convergent. In the second section q-stability is proved from the q-root condition and the 1st q-order condition, and in the third section rth order q-consistency is proved from the q-root condition and the rth q-order condition.

For Chapters 3 and 4 the computed solution $\underline{a}_0$, $\underline{a}_1,\ldots,\underline{a}_N$ is defined by

$$\underline{a}_0 = \underline{a}(0) + \underline{r}_0 \ ,$$

$$\underline{a}_n = S\underline{a}_{n-1} + \underline{\ell}\,\frac{h^p}{p!}\,\Phi(t_n,\underline{a}_{n-1}) + \underline{r}_n, \quad n=1,2,\ldots,N$$

where $\underline{r}_0$ is the starting error and $\underline{r}_1,\underline{r}_2,\ldots,\underline{r}_N$ are errors arising from finite precision arithmetic. These errors are lumped together as the round-off error

$$\underline{R} = (\underline{r}_0,\underline{r}_1,\ldots,\underline{r}_N) \ .$$

We measure $\underline{R}$ with the norm

[illegible]

[illegible]

[illegible]

$$[illegible]$$

$$[illegible]$$

$$[illegible]$$

[illegible]

$$[illegible]$$

$$[illegible]$$

$$|\underline{R}|_S \equiv \max_{0\le n\le N} || \sum_{j=0}^{n} S^{n-j} \underline{r}_j ||_H \quad ,$$

which resembles the norm introduced by Spijker (1971).

Section 3.1 q-stability and q-consistency

Roughly speaking, stability is convergence of order zero; that is, a method is stable if $\max_{0\le n\le N} ||\underline{a}_n||_H$ is uniformly bounded for all h whenever $h^{1-p}||\underline{r}_0||$ is uniformly bounded and $\underline{r}_1=\ldots=\underline{r}_N=\underline{0}$. Such a definition is not very useful for proving theorems or analyzing the effects of round-off error. A much more useful definition is the

Definition of q-stability:

A multivalue method is q-stable if

1. for any Lipschitz problem, there exists a constant k_0 such that

$$||\underline{a}_n - \underline{a}_n^x||_H \le k_0 |\underline{R}|_S$$

where $\underline{a}_0^x, \underline{a}_1^x, \ldots, \underline{a}_N^x$ is the numerical solution with no round-off error ("x" is for "exact").

2. there exists a constant k_S such that

$$||H^{-1} S^n|| \le k_S h^{1-p} \quad . \qquad \square$$

For a q-stable method it follows that

$$||\underline{a}_n-\underline{a}_n^x||_H \leq k_0 k_S h^{1-p} \sum_{n=0}^{N} ||\underline{r}_n|| \;,$$

which is a more conventional definition for stability (see Chartres and Stepleman (1972)).

The usual definition of consistency (Gear (1971: 181)) states that $\tilde{\underline{a}}(t) = \underline{a}(t)+\underline{O}(h^{r+p})$ where

$$\tilde{\underline{a}}(t) \equiv S\underline{a}(t-h) + \tilde{\underline{\ell}} \frac{h^p}{p!} \Phi(t,\underline{a}(t-h)) \;.$$

In other words, the local truncation error arising from the correct solution $\underline{a}(t)$ must be of order h^{r+p}. Although, this definition is satisfactory for multistep methods, it is too strong for multivalue methods. A method which is rth order q-consistent in the usual sense would have to satisfy $D_{ij} = \delta_{ij}$ for $j < \min(i,p)+r$, which is stronger than the corrector condition.

At the beginning of Chapter 1, the convergence of a method is demonstrated by finding an interpretation of the computed values for which the local truncation error is $O(h^2)$ and which differs from the Nordsieck interpretation by $O(h)$. Therefore a method is defined to be q-consistent of order r if there exists an interpretation for which the local truncation error is $O(h^{r+1})$

and which differs from the Nordsieck interpretation by $O(h^r)$. This special interpretation is represented in the following definition by the matrix $\hat{D}$ in the same way that E represents the Nordsieck interpretation and D represents the first r+p terms of the optimal interpretation.

<u>Definition of rth order q-consistency</u>:

Let r be a positive interger. A multivalue method is q-consistent of order r if there exists a matrix $\hat{D}$ such that for any Lipschitz problem with solution $y(t) \varepsilon C^{r+p}[0,1]$, the function $\underline{a}^s(t) = \hat{D}\underline{Y}(t)$ satisfies

1. $||\underline{a}^s(t)-\underline{a}(t)||_H = O(h^r)$

2. $||S^n(\underline{a}^s(0)-\underline{a}(0))||_H = O(h^r)$

3. $\tilde{\underline{a}}^s(t) = \underline{a}^s(t) + \underline{O}(h^{r+p})$

where $\tilde{\underline{a}}^s(t) \equiv S\underline{a}^s(t-h) + \tilde{\underline{\ell}}\,\frac{h^p}{p!}\,\Phi(t,\underline{a}^s(t-h))$. □

Part 2 of the definition has been introduced for the case $q > 1$; it is not necessary if $q = 1$. Part 2 states that the special starting value differs from the correct starting value by $O(h^r)$ when measured by the norm $\max_{0 \le n \le N} ||S^n\cdot||_H$. Part 1 can be deduced from part 2

and is therefore redundant, but it is included to make the definition more meaningful.

The symbol $\hat{D}$ is used in the definition in order to suggest the correspondence between the three parts of q-consistency and the three equations of Lemma 2.2.

In Section 3.3 it is shown that part 3 of rth order q-consistency is equivalent to

$$(3.1) \qquad \Phi(t,\underline{a}^s(t-h)) = y^{(p)}(t) + O(h^r)$$

when $\hat{D} = D$, assuming D exists. If Φ satisfies a Lipschitz condition as in Lemma 3.2 and if part 1 of q-consistency holds, then (3.1) is equivalent to

$$\Phi(t,\underline{a}(t-h)) = y^{(p)}(t) + O(h^r) \ .$$

The following example shows that there are very useful methods for which the usual definition of consistency is inadequate. The 2-step Adams-Moulton method in normal form has the parameters

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \underline{\ell} = \begin{pmatrix} 5/12 \\ 1 \\ 1/2 \end{pmatrix} \ .$$

The truncated optimal interpretation gives us

$$\underline{a}^s(t) = \begin{pmatrix} 1 & 0 & 0 & 1/4 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -3/2 \end{pmatrix} \underline{Y}(t) \quad .$$

The 3rd 1-order condition is satisfied since the first 3 columns of D are identical to those of E. However, the usual definition of 3rd order consistency cannot be satisfied unless $D_{ij} = \delta_{ij}$ for $j < \min(i,1)+3$, which is not the case.

The next example of a $P(EC)^1$ method for second order equations demonstrates that 2-consistency does not imply 1-consistency. For the parameters

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} \quad \text{and} \quad \underline{\ell} = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} \quad ,$$

it can be shown that

$$S = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix} 1 & 0 & -1/6 \\ 0 & 1/2 & 0 \\ 0 & 1/2 & 1 \end{pmatrix} \quad .$$

The 1-root condition and the 1st 2-order condition are satisfied but the 1st 1-order condition is not. Note that the definition of rth order q-consistency depends on q through the use of the norm $||\cdot||_H$.

The following theorem, from Chartres and Stepleman (1972), enables us to establish rth order q-convergence by proving q-stability and rth order q-consistency.

Theorem 3.1:

Consider a method which is q-stable and q-consistent of order r. Then the computed solution for a Lipschitz problem with solution $y(t) \varepsilon C^{r+p}[0,1]$ satisfies

$$||\underline{a}_n - \underline{a}(t_n)||_H \leq k_0 |\underline{R}|_S + O(h^r) \quad ,$$

and therefore the method is q-convergent of order r.

Proof: By the triangle inequality

$$||\underline{a}_n - \underline{a}(t_n)||_H \leq ||\underline{a}_n - \underline{a}_n^x||_H + ||\underline{a}^s(t_n) - \underline{a}_n^x||_H$$

$$+ ||\underline{a}^s(t_n) - \underline{a}_n(t_n)||_H$$

where $\underline{a}^s(t)$ satisfies all three parts of rth order q-consistency. By q-stability

$$||\underline{a}_n - \underline{a}_n^x||_H \leq k_0 |\underline{R}|_S \quad .$$

Notice that $\underline{a}^s(t_0), \underline{a}^s(t_1), \ldots, \underline{a}^s(t_N)$ is the computed solution with

$$\underline{R}^s = (\underline{a}^s(t_0)-\underline{a}(t_0), \underline{a}^s(t_1)-\tilde{\underline{a}}^s(t_1), \ldots,$$

$$\underline{a}^s(t_N)-\tilde{\underline{a}}^s(t_N)) \ ,$$

and thus by q-stability

$$\begin{aligned}||\underline{a}^s(t_n)-\underline{a}_n^x||_H &\le k_0|R^s|_S \\ &\le k_0 \max_{0\le n\le N} ||S^n(\underline{a}^s(t_0)-\underline{a}(t_0))||_H \\ &+ k_0k_Sh^{1-p} \sum_{j=1}^{N} ||\underline{a}^s(t_j)-\tilde{\underline{a}}^s(t_j)||,\end{aligned}$$

and by parts 2 and 3 of rth order q-consistency

$$\begin{aligned}||\underline{a}^s(t_n)-\underline{a}_n^x||_H &= k_0O(h^r) + k_0k_SNO(h^{r+1}) \\ &= O(h^r) \ .\end{aligned}$$

By part 1 of q-consistency

$$||\underline{a}^s(t_n)-\underline{a}(t_n)||_H = O(h^r) \ . \qquad \square$$

Section 3.2 Proving q-stability

The proof of q-stability requires that a method possess two important properties:

1. $||H^{-1}S^n|| \le k_Sh^{1-p}$ for some constant k_S, which is proved from the q-root condition and the 1st q-order condition.

2. for any Lipschitz problem, there exists a constant L such that

$$|\Phi(t,\underline{a}^2)-\Phi(t,\underline{a}^1)| \leq L||\underline{a}^2-\underline{a}^1||_H ,$$

which is proved from the 1st q-order condition.

The first property is a consequence of the following lemma, which examines the size of S^n.

Lemma 3.1:

If a method satisfies the q-root condition and the 1st q-order condition, then the components of S^n are of order

$$\left(\begin{array}{cccc:ccccc}
1 & n & \cdots & n^{p-q-1} & n^{p-q} & \cdots & n^{p-1} & \cdots & n^{p-1} \\
 & 1 & \ddots & \vdots & \vdots & & \vdots & & \vdots \\
 & & \ddots & n & n^2 & \cdots & n^{q+1} & \cdots & n^{q+1} \\
0 & & & 1 & n & \cdots & n^q & \cdots & n^q \\
\hdashline
 & & & & 1 & \cdots & n^{q-1} & \cdots & n^{q-1} \\
 & 0 & & & \vdots & & \vdots & & \vdots \\
 & & & & 1 & \cdots & n^{q-1} & \cdots & n^{q-1}
\end{array}\right).$$

Proof: Partition S as

$$\begin{pmatrix} S_0 & S_1 \\ 0 & S_2 \end{pmatrix}$$

where S_0 is a $(p-q)\times(p-q)$ upper triangular matrix with 1's on the diagonal. Then

$$S^n = \begin{pmatrix} S_0^n & \sum_{\nu=1}^{n} S_0^{n-\nu} S_1 S_2^{\nu-1} \\ \\ 0 & S_2^n \end{pmatrix} .$$

Divide the nonzero portion of S^n into four parts and prove the lemma for each part.

<u>Part I</u>: $0 \leq i \leq j < p-q$. Note that S_{ij} is in the upper right hand corner of a $(j-i+1)\times(j-i+1)$ diagonal block B of S. Since B is upper triangular with 1's on the diagonal $B^n = O(n^{j-i})$. Furthermore $(S^n)_{ij} = (B^n)_{ij}$ which proves that $(S^n)_{ij} = O(n^{j-i})$ as stated in the lemma.

<u>Part II</u>: $i \geq p-q$ and $p-q \leq j < p-1$. The hypotheses of Lemma 2.2 are satisfied with $\hat{D} = D$. Therefore we are entitled to use (2.20), which states that $\underline{d}_j \varepsilon X_j$ where $X_j = (S-I)^{p-1-j} X_{p-1}$ and X_{p-1} is the subspace spanned by the generalized eigenvectors of S corresponding to 1. Thus there exist vectors $\underline{x}_i \varepsilon X_i$ such that

$$\begin{aligned} S^n \underline{\delta}_j &= S^n(\underline{\delta}_j - \underline{d}_j) + S^n \underline{d}_j \\ &= S^n(\underline{\delta}_j - \underline{d}_j) + \underline{d}_j + n\underline{x}_{j-1} + \ldots + \binom{n}{j-p+q} \underline{x}_{p-q} \\ &\quad + \text{some vector in } X_{p-q-1} \, . \end{aligned}$$

By part 2 of the 1st q-order condition

$$(S^n(\underline{\delta}_j - \underline{d}_j))_i = O(n^{j-p+q-1}) \quad \text{for} \quad i \geq p-q .$$

From (2.21)

$$X_{p-q-1} = \langle \underline{\delta}_0, \underline{\delta}_1, \ldots, \underline{\delta}_{p-q-1} \rangle .$$

Hence for $i \geq p-q$

$$(S^n \underline{\delta}_j)_i = O(n^{j-p+q}) \quad .$$

Part III: $i \geq p-q$ and $j \geq p-1$. Recall from the proof of Lemma 2.2 that $(z-1)^p$ is an elementary divisor so that by the q-root condition $\underline{\delta}_j = \underline{x}_{p-1} + \underline{w}$ where $\underline{x}_{p-1} \varepsilon X_{p-1}$ and $S^n \underline{w} = \underline{O}(n^{q-1})$. Then there exist vectors $\underline{x}_i \varepsilon X_i$ such that

$$S^n \underline{\delta}_j = S^n \underline{w} + \underline{x}_{p-1} + n\underline{x}_{p-2} + \ldots + \binom{n}{q-1} \underline{x}_{p-q}$$

$$+ \text{ some vector in } X_{p-q-1} .$$

Therefore for $i \geq p-q$

$$(S^n \underline{\delta}_j)_i = O(n^{q-1}) \; ,$$

and so S^n_2 is as illustrated.

Part IV: $i < p-q$ and $j \geq p-q$. Consider

$$\sum_{\nu=1}^{n} S_0^{n-\nu} S_1 S_2^{\nu-1} = nO(S_0^n)O(S_1)O(S_2^n)$$

where $O(S_0^n)$ and $O(S_2^n)$ are as given in the statement of this lemma, and $O(S_1)$ is a $(p-q)\times(k+q)$ matrix of 1's. □

Corollary:

$||H^{-1}S^n|| \leq k_S h^{1-p}$ for some constant k_S depending only on S. □

The above corollary can be strengthened because actually

$$H^{-1}S^n \underline{\delta}_i = \underline{O}(h^{-\min(i,p-1)}) \ .$$

Therefore there exists a constant k_S' such that

$$|\underline{R}|_S \leq k_S' \sum_{n=0}^{N} ||\underline{r}_n||_h^p$$

where

$$||\underline{a}||_h^p \equiv ||\mathrm{col}(a_0, h^{-1}a_1, \ldots, h^{1-p}a_{p-1}, \ldots, h^{1-p}a_{k+p-1})||.$$

This bound suggests that $\underline{R}$ could be measured by the norm

$$\sum_{n=0}^{N} ||\underline{r}_n||_h^p \quad ,$$

which is independent of the method. Because this norm is cruder than $|\cdot|_S$, it is unsatisfactory for proving theorems in this chapter.

The function Φ resembles the increment function of a one-step method, and therefore it is expected that for a stable method, $\Phi(t,\underline{a})$ should be required to satisfy a Lipschitz condition in $\underline{a}$.

Lemma 3.2:

If a multivalue method satisfies the 1st q-order condition, then for any Lipschitz problem, there exists a constant L such that

$$|\Phi(t,\underline{a}^2) - \Phi(t,a^1)| \leq L||\underline{a}^2-\underline{a}^1||_H \quad .$$

Proof: Let $\underline{e} = \underline{a}^2-\underline{a}^1$ and $\varepsilon = \Phi(t,\underline{a}^2) - \Phi(t,\underline{a}^1)$. Consider three cases.

Case I: The corrector only method. $\varepsilon = \phi^2-\phi^1$ where

$$\phi^\nu = F(t,S\underline{a}^\nu + \underline{\ell}\,\frac{h^p}{p!}\,\phi^\nu) \qquad \nu=1,2.$$

Because f is uniformly Lipschitz,

$$|\varepsilon| \leq \sum_{i=0}^{p-q} L_i\,\frac{i!}{h^i}\,|(S\underline{e})_i + \ell_i\,\frac{h^p}{p!}\,\varepsilon|$$

whence

$$(3.2) \qquad |\varepsilon| \le \frac{\sum_{i=0}^{p-q} i!L_i ||h^{-i}\underline{\delta}_i^T SH|| \, ||\underline{e}||_H}{1 - \sum_{i=0}^{p-q} \frac{i!}{p!} |\ell_i| L_i h^{p-i}} \ .$$

$||h^{-i}\underline{\delta}_i^T SH||$ is bounded because $S_{ij} = 0$ for $j < i \le p-q$. By the definition of corrector only methods, the denominator of (3.2) is no smaller than η. Therefore there exists an L such that $|\varepsilon| \le L||\underline{e}||_H$.

<u>Case II</u>: The $P(EC)^M$ method with $\ell_p \ne 1$. Let $\varepsilon_{(m)} = \phi^2_{(m)} - \phi^1_{(m)}$ and $\underline{e}_{(m)} = \underline{a}^2_{(m)} - \underline{a}^1_{(m)}$ where

$$(3.3) \qquad \phi^\nu_{(0)} = h^{-p}\, p!\,(A\underline{a}^\nu)_p \ ,$$

$$\phi^\nu_{(m)} = (1 - \frac{u_m}{u_M})\phi^\nu_{(0)} + \frac{1}{u_M} \sum_{j=1}^{m} (1-\ell_p)^{m-j} F(t,\underline{a}^\nu_{(j-1)}) \ ,$$

$$\underline{a}^\nu_{(m)} = S\underline{a}^\nu + \underline{\tilde{\ell}} \frac{h^p}{p!} \phi^\nu_{(m)} \ .$$

Using the Lipschitz condition,

$$F(t,\underline{a}^2_{(j-1)}) - F(t,\underline{a}^1_{(j-1)}) = O(||\underline{e}_{(j-1)}||_H)$$

so that

$$(3.4) \qquad \varepsilon_{(m)} = (1 - \frac{u_m}{u_M})\varepsilon_{(0)} + \sum_{j=1}^{m} O(||\underline{e}_{(j-1)}||_H) \ .$$

Also

$$\underline{e}_{(m)} = S\underline{e} + \underline{\tilde{\ell}} \frac{h^p}{p!} \varepsilon_{(m)} \ .$$

Because $S_{ij} = 0$ for $j < \min(i,p-q)$, $||H^{-1}SH||$ is bounded, and so

$$(3.5) \qquad ||\underline{e}_{(m)}||_H = O(||\underline{e}||_H) + O(h^q\varepsilon_{(m)}) \ .$$

By induction on m it can be shown from (3.4) and (3.5) that

$$(3.6) \qquad \varepsilon = \varepsilon_{(M)} = O(||\underline{e}||_H) + O(h^q\varepsilon_{(0)}) \ .$$

Note that $\underline{\delta}_p^T S = (1 - \tilde{\ell}_p)\underline{\delta}_p^T A$ whence

$$A_{pj} = 0 \quad \text{for} \quad j < p-q$$

and

$$||h^{-p}\underline{\delta}_p^T AH|| = h^{-q}||\underline{\delta}_p^T A|| \ .$$

By (3.3)

$$\varepsilon_{(0)} = p!\ h^{-p}\underline{\delta}_p^T A\underline{e}$$

$$= O(h^{-q}||\underline{e}||_H) \ .$$

This result with (3.6) implies that $\varepsilon = O(||\underline{e}||_H)$.

Case III: The $P(EC)^M$ method with $\ell_p = 1$. Using the notation of Case II, there is considerable simplification:

$$\varepsilon_{(m)} = F(t,\underline{a}^2_{(m-1)}) - F(t,\underline{a}^1_{(m-1)}) \ ,$$

$$\underline{e}_{(m-1)} = S\underline{e} + \underline{\ell}\,\frac{h^p}{p!}\,\varepsilon_{(m-1)} \quad .$$

Using the Lipschitz condition,

$$|\varepsilon_{(m)}| \le \sum_{i=0}^{p-q} L_i\,\frac{i!}{h^i}\,|(S\underline{e})_i + \ell_i\,\frac{h^p}{p!}\,\varepsilon_{(m-1)}| \quad .$$

Because $\ell_i = 0$ for $p-c < i \le p-q$, we have

$$\varepsilon_{(m)} = O(||S\underline{e}||_H) + O(h^c\varepsilon_{(m-1)})$$

(3.7) $$= O(||\underline{e}||_H) + O(h^c\varepsilon_{(m-1)}) \quad .$$

By the predictor condition for $r = 1$,

$$A_{pj} = 0 \quad \text{for} \quad j < p-Mc$$

whence

$$||h^{-p}\underline{\delta}_p^T AH|| = O(h^{-Mc}) \quad .$$

By (3.3)

$$\varepsilon_{(0)} = p!\; h^{-p}\underline{\delta}_p^T A\underline{e}$$

(3.8) $$= O(h^{-Mc}||\underline{e}||_H) \quad .$$

Together (3.7) and (3.8) imply $\varepsilon = \varepsilon_{(M)} = O(||\underline{e}||_H)$. □

The following lemma, proved in Miller (1968:21), is the discrete counterpart to Gronwall's inequality. It is useful for solving inequalities which arise in stability proofs.

Lemma 3.3:

Let

$$w_n \leq \gamma + \sum_{j=0}^{n-1} v_j w_j \quad \text{for} \quad n=0,1,2,\ldots$$

where γ, w_n, and v_n are nonnegative. Then

$$w_n \leq \gamma \exp\left(\sum_{j=0}^{n-1} v_j \right) \quad . \qquad \square$$

Theorem 3.2:

A multivalue method which satisfies the q-root condition and the 1st q-order condition is q-stable.

Proof: Consider a Lipschitz problem, and set $\underline{e}_n = \underline{a}_n - \underline{a}_n^x$. Then

$$\underline{e}_n = S\underline{e}_{n-1} + \underline{\tilde{\ell}} \frac{h^p}{p!} \varepsilon_n + \underline{r}_n$$

where $\varepsilon_n = \Phi(t_n, \underline{a}_{n-1}) - \Phi(t_n, \underline{a}_{n-1}^x)$, and by Lemma 3.2 there exists a constant L such that

$$(3.9) \qquad |\varepsilon_n| \leq L||\underline{e}_{n-1}||_H \quad .$$

Solving the difference equation gives

$$\underline{e}_n = S^n \underline{r}_0 + \sum_{j=1}^{n} S^{n-j} \left(\tilde{\underline{\ell}} \frac{h^p}{p!} \varepsilon_j + \underline{r}_j\right) .$$

Therefore

$$||\underline{e}_n||_H \le \frac{h^p}{p!} \sum_{j=1}^{n} |\varepsilon_j| \, ||S^{n-j}\tilde{\underline{\ell}}||_H + ||\sum_{j=0}^{n} S^{n-j}\underline{r}_j||_H .$$

From the corollary to Lemma 3.1 it follows that

$$||S^{n-j}\tilde{\underline{\ell}}||_H \le k_S h^{1-p} ||\tilde{\underline{\ell}}|| .$$

This inequality and (3.9) imply that there is a constant k_1 such that

$$||\underline{e}_n||_H \le hk_1 \sum_{j=1}^{n} ||\underline{e}_{j-1}||_H + |\underline{R}|_S .$$

Apply Lemma 3.3 with $w_n = ||\underline{e}_n||_H$, $\gamma = |\underline{R}|_S$, and $v_n = hk_1$ to get

$$||\underline{e}_n||_H \le |\underline{R}|_S \exp(nhk_1) \le |\underline{R}|_S \exp(k_1) .$$

Together with the corollary to Lemma 3.1, this proves q-stability. □

Section 3.3 Proving rth Order q-consistency

The proof of rth order q-consistency assumes the rth q-order condition, and so the choice of $\underline{a}^s(t) = D\underline{Y}(t)$

automatically satisfies $||\underline{a}^s(t)-\underline{a}(t)||_H = O(h^r)$. To show that $||S^n(\underline{a}^s(0)-\underline{a}(0)||_H = O(h^r)$, we first prove the following lemma, which has the q-root condition as an additional hypothesis.

Lemma 3.4:

If a multivalue method satisfies the q-root condition and the rth q-order condition, then

$$(S^n(D-E))_{ij} = O(n^{j-r-\min(i,p-q)}) \ .$$

Proof: Let $\underline{u}_j = \underline{d}_j-\underline{\delta}_j = (D-E)\underline{\varepsilon}_j$, and consider four sets of values for j.

Case I: $j < r$. Then $\underline{u}_j = \underline{0}$.

Case II: $r \leq j < r+p-q$. $(\underline{u}_j)_i = 0$ for $i > j-r$. From Lemma 3.1 it can be seen that

$$(S^n\underline{u}_j)_i = O(n^{j-r-i}) \ .$$

Case III: $r+p-q \leq j < r+p-1$. By the rth q-order condition

$$(S^n\underline{u}_j)_i = O(n^{j-r-p+q}) \quad \text{for} \quad i \geq p-q \ .$$

Partition S as in Lemma 3.1 and likewise partition

$$\underline{u}_j = \begin{pmatrix} \underline{u}^1 \\ \underline{u}^2 \end{pmatrix} .$$

Thus

$$S^n\underline{u}_j = \begin{bmatrix} S_0^n\underline{u}^1 + \sum_{\nu=1}^{n} S_0^{n-\nu}S_1S_2^{\nu-1}\underline{u}^2 \\ S_2^n\underline{u}^2 \end{bmatrix}$$

where $S_2^n\underline{u}^2 = \underline{O}(n^{j-r-p+q})$. The size of S_0^n is known from Lemma 3.1, and so it is easy to show that

$$(S^n\underline{u}_j)_i = O(n^{j-r-\min(i,p-q)}) .$$

<u>Case IV</u>: $j = r+p-1$. From Lemma 3.1, it is obvious that

$$S^n\underline{u}_{r+p-1} = \mathrm{col}(O(n^{p-1}),\ldots,O(n^{q-1}),\ldots,O(n^{q-1})) . \qquad \square$$

<u>Theorem 3.3</u>:

If a multivalue method satisfies the q-root condition and the rth q-order condition, then it is q-consistent of order r.

<u>Proof</u>: With $\underline{a}^s(t) = D\underline{Y}(t)$ the first requirement of q-consistency follows from part 1 of the rth q-order condition. Applying Lemma 3.4 with h^{-1} substituted for n gives

$$[S^n(\underline{a}^s(0)-\underline{a}(0))]_i = (S^n(D-E)\underline{Y}(0))_i$$

$$= \sum_{j=0}^{r+p-1} (S^n(D-E))_{ij} O(h^j)$$

$$= O(h^{r+\min(i,p-q)})$$

whence

$$||S^n(\underline{a}^s(0)-\underline{a}(0))||_H = O(h^r) \quad ,$$

thus proving the second part of q-consistency. Recall that

$$(3.10) \qquad P\underline{Y}(t-h) = \underline{Y}(t) + \underline{O}(h^{r+p}) \quad .$$

Use this and Lemma 2.1 to deduce that

$$S\underline{a}^s(t-h) = SD\underline{Y}(t-h)$$

$$= D\underline{Y}(t) - \underline{\tilde{\ell}}\, \underline{\varepsilon}_p^T \underline{Y}(t) + \underline{O}(h^{r+p})$$

$$(3.11) \qquad = \underline{a}^s(t) - \underline{\tilde{\ell}} \frac{h^p}{p!} y^{(p)}(t) + \underline{O}(h^{r+p}) \quad .$$

By definition

$$(3.12) \qquad \underline{\tilde{a}}^s(t) = S\underline{a}^s(t-h) + \underline{\tilde{\ell}} \frac{h^p}{p!} \Phi(t,\underline{a}^s(t-h)) \quad ,$$

and so we need only show that

$$\Phi(t,\underline{a}^s(t-h)) = y^{(p)}(t) + O(h^r) \quad .$$

Case I: The corrector only method. We have $\Phi(t,\underline{a}^s(t-h)) = \phi$ where $\phi = F(t,\underline{\tilde{a}}^s(t))$. Also $y^{(p)}(t) = F(t,\underline{a}(t))$, and by using the Lipschitz condition,

$$(3.13) \qquad |\phi - y^{(p)}(t)| \le \sum_{i=0}^{p-q} L_i \frac{i!}{h^i} |\tilde{a}_i^s(t) - a_i(t)| \ .$$

From (3.11) and (3.12)

$$\underline{\tilde{a}}^s(t) = \underline{a}^s(t) + \underline{\ell} \frac{h^p}{p!} (\phi - y^{(p)}(t)) + O(h^{r+p}) \ .$$

For $i \le p-q$

$$\tilde{a}_i^s(t) - a_i(t) = \ell_i \frac{h^p}{p!} (\phi - y^{(p)}(t)) + O(h^{i+r}) \ .$$

Substitute this into (3.13):

$$|\phi - y^{(p)}(t)| \le \sum_{i=0}^{p-q} L_i \frac{i!}{h^i} |\ell_i \frac{h^p}{p!} (\phi - y^{(p)}(t)) + O(h^{i+r})| \ ,$$

which gives

$$(1 - \frac{h^p}{p!} \sum_{i=0}^{p-q} L_i \frac{i!}{h^i} |\ell_i|) |\phi - y^{(p)}(t)| = O(h^r) \ .$$

Since the first factor is no smaller than η,

$$\phi = y^{(p)}(t) + O(h^r) \ .$$

Case II: The $P(EC)^M$ method with $\ell_p \ne 1$. For this case $\Phi(t,\underline{a}^s(t-h)) = \phi_{(M)}$ where

$$(3.14) \qquad \phi_{(0)} = h^{-p}\, p!\, (A\underline{a}^s(t-h))_p \ ,$$

$$(3.15)\qquad \phi_{(m)} = (1 - \frac{u_m}{u_M})\ \phi_{(0)} + \frac{1}{u_M} \sum_{j=1}^{m} (1-\ell_p)^{m-j} F(t,\underline{a}_{(j-1)}),$$

$$\underline{a}_{(m)} = S\underline{a}^s(t-h) + \underline{\tilde{\ell}}\ \frac{h^p}{p!}\ \phi_{(m)} \quad .$$

Let $\varepsilon_{(m)} = \phi_{(m)} - y^{(p)}(t)$ and $\underline{e}_{(m)} = \underline{a}_{(m)} - \underline{a}(t)$. Note that

$$(3.16)\qquad y^{(p)}(t) = (1 - \frac{u_m}{u_M}) y^{(p)}(t) + \frac{1}{u_M} \sum_{j=1}^{m} (1-\ell_p)^{m-j} F(t,\underline{a}(t)).$$

Subtract (3.16) from (3.15), and use the Lipschitz property to get

$$(3.17)\qquad \varepsilon_{(m)} = (1 - \frac{u_m}{u_M})\ \varepsilon_{(0)} + \sum_{j=1}^{m} O(||\underline{e}_{(j-1)}||_H) \quad .$$

Also by (3.11)

$$\begin{aligned}\underline{e}_{(m)} &= S\underline{a}^s(t-h) + \underline{\tilde{\ell}}\ \frac{h^p}{p!}\ \phi_{(m)} - \underline{a}(t)\\ &= \underline{a}^s(t) + \underline{O}(h^{r+p}) + \underline{\tilde{\ell}}\ \frac{h^p}{p!}\ \varepsilon_{(m)} - \underline{a}(t)\\ &= \underline{\tilde{\ell}}\ \frac{h^p}{p!}\ \varepsilon_{(m)} + H\underline{O}(h^r)\end{aligned}$$

and so

$$(3.18)\qquad ||\underline{e}_{(m)}||_H = O(h^r) + O(h^q \varepsilon_{(m)}) \quad .$$

Write Lemma 2.1 as

$$(I-\underline{\tilde{\ell}}\ \underline{\delta}_p^T)AD = (I-\underline{\tilde{\ell}}\ \underline{\delta}_p^T)EP + (D-E)P$$

whence

$$(3.19) \qquad \underline{\delta}_p^T AD = \underline{\varepsilon}_p^T P + (1-\tilde{\ell}_p)^{-1} \, \underline{\delta}_p^T (D-E)P$$

so that using (3.10)

$$\underline{\delta}_p^T AD\underline{Y}(t-h) = Y_p(t) + O(h^{r+p-q}) \ .$$

It follows from (3.14) that

$$\begin{aligned} \varepsilon_{(0)} &= h^{-p} \, p! \, \underline{\delta}_p^T AD\underline{Y}(t-h) - y^{(p)}(t) \\ &= O(h^{r-q}) \ . \end{aligned}$$

This result with (3.17) and (3.18) implies that

$$(3.20) \qquad \varepsilon_{(m)} = (1 - \frac{u_m}{u_M}) \varepsilon_{(0)} + O(h^r) = O(h^{r-q}) \ ,$$

and consequently $\phi_{(M)} - y^{(p)}(t) = \varepsilon_{(M)} = O(h^r)$.

<u>Case III</u>: The $P(EC)^M$ method with $\ell_p = 1$. The expression for $\varepsilon_{(m)}$ is very simple:

$$\varepsilon_{(m)} = F(t,\underline{a}_{(m-1)}) - F(t,\underline{a}(t)) \ ,$$

and as in case II

$$\underline{e}_{(m-1)} = H\underline{O}(h^r) + \underline{\ell} \, \frac{h^p}{p!} \, \varepsilon_{(m-1)} \ .$$

The Lipschitz property enables us to write

$$|\varepsilon_{(m)}| \le \sum_{i=0}^{p-q} L_i \frac{i!}{h^i} |O(h^{i+r}) + \ell_i \frac{h^p}{p!} \varepsilon_{(m-1)}| ,$$

and because $\ell_i = 0$ for $p-c < i \le p-q$,

$$\varepsilon_{(m)} = O(h^r) + O(h^c \varepsilon_{(m-1)}) .$$

Hence

$$(3.21) \quad \varepsilon_{(m)} = O(h^r) + O(h^{mc} \varepsilon_{(0)}) .$$

Part 3 of the rth q-order condition states that

$$(AE)_{pj} = A_{pj} = P_{pj} \quad \text{for} \quad j < r+p-Mc .$$

Thus $A_{pi} = 0$ for $i < p-Mc$. Let $j < r+p-Mc$ so that $(D-E)_{ij} = 0$ for $i > j-r$. Since $j-r < p-Mc$,

$$(A(D-E))_{pj} = 0 \qquad \text{for} \quad j < r+p-Mc ,$$

and hence

$$(AD)_{pj} = P_{pj} \qquad \text{for} \quad j < r+p-Mc ,$$

which implies

$$(AD\underline{Y}(t-h))_p = Y_p(t) + O(h^{r+p-Mc}) .$$

By (3.14)

$$\varepsilon_{(0)} = h^{-p}\, p!\, (AD\underline{Y}(t-h))_p - y^{(p)}(t)$$

$$= O(h^{r-Mc}) \quad .$$

This result with (3.21) implies that $\varepsilon_{(M)} = O(h^r)$. □

CHAPTER 4

ASYMPTOTIC BEHAVIOUR OF THE ERROR

In this chapter asymptotic estimates of the global error are found for corrector only methods and $P(EC)^M$ methods. It is shown that the global error $\underline{e}_n$ defined by $\underline{e}_n = \underline{a}_n - \underline{a}(t_n)$ satisfies an equation of the form

$$H^{-1}\underline{e}_n = h^r \underline{e}(t_n) + \underline{O}(h^{r+1})$$

provided that $f(t,y,\ldots,y^{(p-q)})$ is sufficiently smooth and there is no round-off error. Here the function $\underline{e}(t)$ is the magnified error function, which depends on t alone.

The principal results of this chapter are contained in Theorem 4.1. Because the statement of this theorem is long and the proof is tedious, definitions of various quantities appearing in the theorem are presented here. The definitions of the matrices D and E are changed for this chapter only. Instead of D and E being $(k+p)\times(r+p)$, they are $(k+p)\times(r+p+1)$, and similarly, $\underline{\varepsilon}_i^T$ has r+p+1 components for this chapter. Furthermore, the Pascal triangle matrix P is redefined to be a $(r+p+1)\times(r+p+1)$ matrix. Lemma 2.1 clearly holds for these enlarged matrices:

[illegible]

[illegible]

[illegible]

$$[illegible]$$

[illegible]

[illegible]

$$(4.1) \qquad SD = DP - \underline{\tilde{\ell}}\,\underline{\varepsilon}_p^T P \; .$$

Also define

$$\underline{Y}(t) = \mathrm{col}(y(t), hy'(t), \ldots, h^{r+p} y^{(r+p)}(t)/(r+p)!)$$

and

$$J_i(t) = \frac{\partial f}{\partial y^{(i)}}(t, y(t), \ldots, y^{(p-q)}(t)), \quad i=0,1,\ldots,p-q \; ,$$

assuming $\partial f/\partial y^{(i)}$ exists.

The following lemma is used in the proof of Theorem 4.1.

Lemma 4.1:

Let $\partial f/\partial y^{(i)}$ have continuous partial derivatives for $t \varepsilon [0,1]$, and let $\partial^2 f/\partial y^{(i)}\partial y^{(j)}$ be bounded for $t \varepsilon [0,1]$. Then

$$(4.2) \qquad F(t,\underline{a}) = y^{(p)}(t) + \sum_{i=0}^{p-q} J_i(t)(h^{-i} i! a_i - y^{(i)}(t))' + \left(\sum_{i=0}^{p-q} O(h^{-i} i! a_i - y^{(i)}(t))\right)^2 \; .$$

Proof: By two applications of the mean value theorem

$$F(t,\underline{a}) = F(t,\underline{a}(t)) + \sum_{i=0}^{p-q} \frac{\partial F(t,\underline{a}(t))}{\partial a_i}(a_i - a_i(t))$$

$$+ \sum_{i=0}^{p-q} \sum_{j=0}^{p-q} \frac{\partial^2 F(t, \underline{a}(t) + \theta_i[\underline{a} - \underline{a}(t)])}{\partial a_i \partial a_j}(a_i - a_i(t))(a_j - a_j(t))$$

where $\theta_i \varepsilon (0,1)$ depends on t and $\underline{a}$. The lemma follows by noting that $\partial/\partial a_i = h^i i! \frac{\partial}{\partial y^{(i)}}$. □

The following theorem gives the asymptotic behaviour of the numerical solution for a special class of starting values.

Theorem 4.1:

Consider an rth order q-convergent multivalue method. For a $P(EC)^M$ method it is required that if $\ell_p = 1$ and $0 \leq r+p-Mc \leq p-c$ then $A_{p,r+p-Mc} = 0$. Let $y(t) \varepsilon C^{r+p+1}[0,1]$ be the solution of a Lipschitz problem for which f satisfies the hypotheses of Lemma 4.1. Let the starting value

$$\underline{a}_0 = D\underline{Y}(0) + \sum_{i=0}^{p-1} \underline{d}_i \frac{h^{i+r}}{i!} \delta_0^{(i)} + \underline{r}_0$$

where $\delta_0, \delta_0', \ldots, \delta_0^{(p-1)}$ are freely chosen. Then

$$||\underline{a}_n - \underline{a}^1(t_n)||_H \leq k_0 |\underline{R}|_S + O(h^{r+1})$$

where

$$\underline{a}^1(t) = D\underline{Y}(t) + \sum_{i=0}^{p-1} \underline{d}_i \frac{h^{i+r}}{i!} \delta^{(i)}(t) \quad .$$

The function $\delta(t)$ solves the initial value problem

$$\delta^{(i)}(0) = \delta_0^{(i)} \qquad \text{for} \quad i=0,1,\ldots,p-1 ,$$

$$\delta^{(p)}(t) = \sum_{i=0}^{p-q} J_i(t)\left(\delta^{(i)}(t) + \frac{i!}{(i+r)!} D_{i,i+r} y^{(i+r)}(t)\right) + x(t)$$

where $x(t)$ is specified for each method as follows:

1. for a corrector only method and for a $P(EC)^M$ method with $\ell_p = 1$ and $r+p-Mc \leq p-c$

$$x(t) = 0 .$$

2. for a $P(EC)^M$ method with $\ell_p \neq 1$

$$x(t) = \frac{(p-q)!}{(r+p-q)!} \frac{\ell_{p-q}}{\ell_p} \left(1 - \frac{M(1-\ell_p)^{M-1}}{u_M}\right)(D-E)_{p,r+p-q} \times$$

$$J_{p-q}(t) y^{(r+p-q)}(t) .$$

3. for a $P(EC)^M$ method with $\ell_p = 1$ and $r+p-Mc > p-c$

$$x(t) = \left[\ell_{p-c} \frac{(p-c)!}{p!} J_{p-c}(t)\right]^M \frac{p!}{(r+p-Mc)!} (AD-EP)_{p,r+p-Mc} \times$$

$$y^{(r+p-Mc)}(t) .$$

<u>Proof</u>: Define

$$\underline{\Delta}(t) = \mathrm{col}(\delta(t), h\delta'(t), \ldots, \frac{h^p}{p!}\delta^{(p)}(t), 0, \ldots, 0)$$

and $\underline{a}^s(t) = D(\underline{Y}(t) + h^r \underline{\Delta}(t))$. Since $\underline{a}^1(t) = \underline{a}^s(t) + O(h^{r+p})$,

it is sufficient to consider $||\underline{a}_n-\underline{a}^s(t_n)||_H$. By the triangle inequality

$$||\underline{a}_n-\underline{a}^s(t_n)||_H \le ||\underline{a}_n-\underline{a}_n^x||_H + ||\underline{a}^s(t_n)-\underline{a}_n^x||_H \ .$$

By q-stability a bound on the first term is given by

$$||\underline{a}_n-\underline{a}_n^x||_H \le k_0|\underline{R}|_S \ .$$

A bound on the second term is found by using q-stability and the fact that $\underline{a}^s(t_0),\underline{a}^s(t_1),\dots,\underline{a}^s(t_N)$ is the computed solution when

$$\underline{R}^s = (\underline{a}^s(t_0)-\underline{a}^1(t_0),\underline{a}^s(t_1)-\tilde{\underline{a}}^s(t_1),\dots,\underline{a}^s(t_N)-\tilde{\underline{a}}^s(t_N))$$

where

$$(4.3) \qquad \tilde{\underline{a}}^s(t) = S\underline{a}^s(t-h) + \underline{\tilde{\ell}}\,\frac{h^p}{p!}\,\Phi(t,\underline{a}^s(t-h)) \ .$$

Thus

$$||\underline{a}^s(t_n)-\underline{a}_n^x||_H \le k_0k_Sh^{1-p}||\underline{a}^s(t_0)-\underline{a}^1(t_0)|| + k_0k_Sh^{1-p}\sum_{j=1}^{N}||\underline{a}^s(t_j)-\tilde{\underline{a}}^s(t_j)||.$$

Therefore we need only show that

$$\tilde{\underline{a}}^s(t) = \underline{a}^s(t) + \underline{O}(h^{r+p+1}) \ .$$

Since $\delta(t) \varepsilon C^{p+1}[0,1]$,

$$(4.4) \qquad P(\underline{Y}(t-h)+h^r\underline{\Delta}(t-h)) = \underline{Y}(t)+h^r\underline{\Delta}(t) + \underline{O}(h^{r+p+1}) \ .$$

Use this and (4.1) to deduce that

$$(4.5) \qquad S\underline{a}^s(t-h) = \underline{a}^s(t) - \underline{\tilde{\ell}}\,\frac{h^p}{p!}\,(y^{(p)}(t)+h^r\delta^{(p)}(t))+\underline{O}(h^{r+p+1}) .$$

After substituting (4.5) into (4.3), it becomes clear that we need only show that

$$(4.6) \qquad \Phi(t,\underline{a}^s(t-h)) = y^{(p)}(t) + h^r\delta^{(p)}(t) + O(h^{r+1}) \ .$$

<u>Case I</u>: The corrector only method. In this case

$$\Phi(t,\underline{a}^s(t-h)) = \phi = F(t,\underline{\tilde{a}}^s(t)) \ .$$

From (4.3) and (4.5) it follows that

$$\underline{\tilde{a}}^s(t) = \underline{a}^s(t) + \underline{\tilde{\ell}}\,\frac{h^p}{p!}\,(\phi-y^{(p)}(t)) + \underline{O}(h^{r+p}) \ .$$

In the proof of Theorem 3.3 it was shown that

$$\phi = y^{(p)}(t) + O(h^r)$$

and thus for $i \le p-q$

$$h^{-i}i!\tilde{a}_i^s(t) = y^{(i)}(t) + h^r w_i(t) + O(h^{r+1})$$

where

where

$$w_i(t) \equiv \delta^{(i)}(t) + \frac{i!}{(i+r)!} D_{i,i+r} y^{(i+r)}(t) .$$

Apply (4.2) with $\underline{a} = \underline{\tilde{a}}^s(t)$ to get

$$\begin{aligned} \phi &= y^{(p)}(t) + \sum_{i=0}^{p-q} J_i(t)(h^r w_i(t) + O(h^{r+1})) + O(h^{2r}) \\ &= y^{(p)}(t) + h^r \delta^{(p)}(t) + O(h^{r+1}) \end{aligned}$$

thus showing (4.6).

<u>Case II</u>: The $P(EC)^M$ method with $\ell_p \neq 1$. In this case $\Phi(t,\underline{a}^s(t-h)) = \phi_{(M)}$ where

$$\phi_{(0)} = h^{-p} p! (A\underline{a}^s(t-h))_p ,$$

$$\phi_{(m)} = (1 - \frac{u_m}{u_M})\phi_{(0)} + \frac{1}{u_M} \sum_{j=1}^{m} (1-\ell_p)^{m-j} F(t,\underline{a}_{(j-1)}) ,$$

$$\underline{a}_{(m)} = S\underline{a}^s(t-h) + \underline{\tilde{\ell}} \frac{h^p}{p!} \phi_{(m)} .$$

Let $\varepsilon_{(m)} = \phi_{(m)} - y^{(p)}(t)$. By (4.5),

$$\begin{aligned} \underline{a}_{(m)} &= \underline{a}^s(t) + \underline{\tilde{\ell}} \frac{h^p}{p!} (\phi_{(m)} - y^{(p)}(t)) + \underline{O}(h^{r+p}) \\ &= \underline{a}^s(t) + \underline{\tilde{\ell}} \frac{h^p}{p!} \varepsilon_{(m)} + \underline{O}(h^{r+p}) \end{aligned}$$

so that for $i \leq p-q$

$$(4.7)\qquad h^{-i}i!(\underline{a}_{(m)})_i = y^{(i)}(t)+h^r w_i(t)+h^{p-i}\,\frac{i!\tilde{\ell}_i}{p!}\,\varepsilon_{(m)} + O(h^{r+1}) \; .$$

From (3.20)

$$(4.8)\qquad \varepsilon_{(m)} = (1-\frac{u_m}{u_M})\varepsilon_{(0)} + O(h^r) = O(h^{r-q}) \; .$$

Apply (4.2) with $\underline{a} = \underline{a}_{(m)}$ to get

$$F(t,\underline{a}_{(m)}) = y^{(p)}(t) + h^r \sum_{i=0}^{p-q} J_i(t)w_i(t) + J_{p-q}(t)\,\frac{(p-q)!}{p!}\,h^q\tilde{\ell}_{p-q}\varepsilon_{(m)} + O(h^{r+1}).$$

Use this result to obtain

$$\phi_{(M)} = \frac{1}{u_M}\sum_{j=1}^{M}(1-\ell_p)^{M-j}F(t,\underline{a}_{(j-1)})$$

$$(4.9)\qquad = y^{(p)}(t) + h^r\sum_{i=0}^{p-q} J_i(t)w_i(t) + J_{p-q}(t)\,\frac{(p-q)!}{p!}\,h^q\,\frac{\tilde{\ell}_{p-q}}{u_M}\sum_{j=1}^{M}(1-\ell_p)^{M-j}\varepsilon_{(j-1)} + O(h^{r+1}).$$

Recall (3.19):

$$\underline{\delta}_p^T AD = \underline{\varepsilon}_p^T P + (1-\tilde{\ell}_p)^{-1}\,\underline{\delta}_p^T(D-E)P \; .$$

Using (4.4)

$$\varepsilon_{(0)} = h^{-p}p!(A\underline{a}^s(t-h))_p - y^{(p)}(t)$$

$$= h^{-p}p!\underline{\delta}_p^T AD(\underline{Y}(t-h) + h^r\underline{\Delta}(t-h)) - y^{(p)}(t)$$

$$= h^{-p} p!\,(\underline{\varepsilon}_p^T + (1-\tilde{\ell}_p)^{-1}\underline{\delta}_p^T (D-E))(\underline{Y}(t) + h^r \underline{\Delta}(t)) - y^{(p)}(t) + O(h^{r+p+1})$$

$$= h^{r-q} \frac{p!\,(D-E)_{p,r+p-q}}{(r+p-q)!\,(1-\tilde{\ell}_p)} y^{(r+p-q)}(t) + O(h^{r+q+1}) \ .$$

Substitute this into (4.8), and then put the result into (4.9). After considerable algebraic manipulation, it turns out that

$$\phi_{(M)} = y^{(p)}(t) + h^r \delta^{(p)}(t) + O(h^{r+1}) \ .$$

Case III: The $P(EC)^M$ method with $\ell_p = 1$ and $r+p-Mc > p-c$. In the proof of Theorem 3.3, (3.22) clearly holds for the augmented D:

$$(AD)_{pj} = P_{pj} \qquad \text{for} \quad j < r+p-Mc$$

whence

$$(\underline{\delta}_p^T AD)_j = ((\underline{\varepsilon}_p^T + \alpha \underline{\varepsilon}_{r+p-Mc}^T) P)_j \qquad \text{for } j \le r+p-Mc$$

where

$$\alpha = (AD-EP)_{p,r+p-Mc} \ .$$

Therefore

$$\underline{\delta}_p^T AD\underline{Y}(t-h) = \frac{h^p}{p!} y^{(p)}(t) + \alpha \frac{h^{r+p-Mc}}{(r+p-Mc)!} y^{(r+p-Mc)}(t) + O(h^{r+p-Mc+1}),$$

and from (4.4)

$$\varepsilon_{(0)} = h^{-p} p!(AD\underline{Y}(t-h))_p - y^{(p)}(t)$$

$$(4.10) \qquad = h^{r-Mc} \frac{\alpha p!}{(r+p-Mc)!} y^{(r+p-Mc)}(t) + O(h^{r-Mc+1}) \ .$$

From (3.21), we know that for $m \leq M$

$$\varepsilon_{(m)} = O(h^r) + O(h^{mc}\varepsilon_{(0)}) = O(h^{r+(m-M)c}) \ .$$

Thus for $i \leq p-q$

$$h^{p-i} \frac{i!\ell_i}{p!} \varepsilon_{(m-1)} = \begin{cases} O(h^{r+(m-M)c+1}) & \text{if } i < p-c \\ h^c \dfrac{(p-c)!\ell_{p-c}}{p!} \varepsilon_{(m-1)} & \text{if } i = p-c \\ 0 & \text{if } i > p-c \ . \end{cases}$$

It follows from (4.7) with $1 \leq m \leq M$ that

$$h^{-i} i!(\underline{a}_{(m-1)})_i = y^{(i)}(t) + h^r w_i(t) + O(h^{r+(m-M)c+1})$$

$$+ \begin{cases} h^c \dfrac{(p-c)!}{p!} \ell_{p-c}\varepsilon_{(m-1)} & \text{if } i = p-c \\ 0 & \text{otherwise} \ . \end{cases}$$

Apply (4.2) with $\underline{a} = \underline{a}_{(m-1)}$ to get

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

$$\varepsilon_{(m)} = F(t,\underline{a}_{(m-1)}) - y^{(p)}(t)$$

$$= h^r \sum_{i=0}^{p-q} J_i(t)w_i(t) + h^c J_{p-c}(t)\frac{(p-c)!}{p!}\ell_{p-c}\varepsilon_{(m-1)}$$

$$+ O(h^{r+(m-M)c+1}) + O[h^{2(r+(m-M)c)}] .$$

$$(4.11) \qquad = h^r \sum_{i=0}^{p-q} J_i(t)w_i(t) + h^c J_{p-c}(t)\frac{(p-c)!}{p!}\ell_{p-c}\varepsilon_{(m-1)}$$

$$+ O(h^{r+(m-M)c+1})$$

because $r+(m-M)c \geq r+(1-M)c > 0$. Notice that if $m < M$ then the first term of (4.11) can be absorbed into $O(h^{r+(m-M)c+1})$. Therefore using (4.10) and (4.11) to solve for $\varepsilon_{(M)}$ gives

$$\varepsilon_{(M)} = h^r \delta^{(p)}(t) + O(h^{r+1}) ,$$

which proves (4.6).

Case IV: The $P(EC)^M$ method with $\ell_p = 1$ and $r+p-Mc \leq p-c$. Because of the restriction on the method

$$A_{pj} = 0 \qquad \text{for} \qquad j \leq r+p-Mc ,$$

which implies

$$\varepsilon_{(0)} = h^{-p}p!(A\underline{a}^s(t-h))_p - y^{(p)}(t)$$

$$= O(h^{r-Mc+1}) ,$$

and so $\varepsilon_{(M-1)} = O(h^{r-c+1})$. Thus for $i \leq p-q$

$$h^{p-i}\,\frac{i!}{p!}\,\ell_i\,\varepsilon_{(M-1)} = O(h^{r+1}) \ .$$

It follows from (4.7) that

$$h^{-i}\,i!\,(\underline{a}_{(M-1)})_i = y^{(i)}(t) + h^r w_i(t) + O(h^{r+1})$$

from which it is straightforward to get

$$\varepsilon_{(M)} = h^r \delta^{(p)}(t) + O(h^{r+1}) \ . \qquad \square$$

<u>Corollary</u>:

Assume that the method is strongly stable and that $f = f(t)$. Let $0 < \varepsilon \le 1$ be given. Then for $\underline{a}_0 = \underline{a}(0)$ and $\underline{r}_1 = \ldots = \underline{r}_N = \underline{0}$, we have

$$\underline{a}_n = \underline{a}^1(t_n) + H\underline{O}(h^{r+1}) \text{ uniformly for } t_n \varepsilon [\varepsilon,1]$$

where

$$\delta_0^{(i)} = \frac{-i!}{(i+r)!}\, y^{(i+r)}(0) \lim_{n\to\infty} n^{-i}(S^n(D-E))_{0,i+r} \ .$$

<u>Proof</u>: For the differential equation $y^{(p)} = f(t)$ it is easily seen that

$$\underline{a}_n = \underline{a}_n^x + S^n(\underline{a}(0) - \underline{a}^1(0)) \ .$$

By Theorem 4.1, $\underline{a}_n^x = \underline{a}^1(t_n) + H\underline{O}(h^{r+1})$, and hence we need only show that

$$(4.12) \quad S^n(\underline{a}(0) - \underline{a}^1(0)) = H\underline{O}(h^{r+1}) \quad \text{for} \quad n \geq \varepsilon N.$$

To verify (4.12), it is sufficient to show for $i = 0,1,\ldots,p-1$

$$(4.13) \quad S^n[(D-E)\underline{\varepsilon}_{i+r} - \underline{d}_i \lim_{n\to\infty} n^{-i}(S^n(D-E))_{0,i+r}] = H\underline{O}(h^{1-i})$$

for $n \geq \varepsilon N$. Let X_i be defined as in the proof of Lemma 2.2, and let W be the space spanned by all generalized eigenvectors corresponding to eigenvalues other than 1. Because the method is strongly stable, dim $X_j = j+1$. Furthermore, if $\underline{w} \varepsilon W$ then $S^n\underline{w} = \underline{O}(N^{-1})$ for $n \geq \varepsilon N$. By the rth q-order condition, if $i < p-q$ then

$$(D-E)\underline{\varepsilon}_{i+r} \ \varepsilon <\underline{\delta}_0,\underline{\delta}_1,\ldots,\underline{\delta}_i> \ = X_i = <\underline{d}_0,\underline{d}_1,\ldots,\underline{d}_i> .$$

The coefficient of $\underline{d}_i$ in (4.13) has been chosen so that the expression in square brackets is a vector in X_{i-1}. With this information it is straightforward to verify (4.13). Using Lemma 3.4, it is not difficult to show that if $i \geq p-q$ then $(D-E)\underline{\varepsilon}_{i+r} \varepsilon X_i + W$. The coefficient of $\underline{d}_i$ has been chosen so that the expression in square brackets is in $X_{i-1} + W$. Once again, the verification of (4.13) is straightforward. □

Results of Henrici (1962:255) cause us to believe that this corollary is also valid for $f = f(t,y,\ldots,y^{(p-q)})$.

CHAPTER 5

VARIABLE STEPSIZE

Here let us consider the general grid

$$0 = t_0 < t_1 < \ldots < t_N = 1.$$

To measure the fineness of the grid, define

$$h = \max_{1 \le n \le N} h_n$$

where $h_n \equiv t_n - t_{n-1}$. For this chapter the dependence of various quantities on the stepsize h_n is made explicit in the notation.

Let $\alpha_n = h_n/h_{n-1}$ where $h_0 = h_1$, and define $C_n = \mathrm{diag}(1, \alpha_n, \ldots, \alpha_n^{k+p-1})$. Notice that we have $\underline{a}(t_{n-1};h_n) = C_n\underline{a}(t_{n-1};h_{n-1})$. This equality is the rationale for the interpolatory technique of varying stepsize.

<u>Definition of variable stepsize multivalue method</u>:

$$\underline{a}_0 = \underline{a}(0;h_0) + \underline{r}_0 \ ,$$

$$\underline{a}_n = SC_n\underline{a}_{n-1} + \underline{\tilde{\ell}}\,\frac{h^p}{p!}\,\Phi(t_n, C_n\underline{a}_{n-1}; h_n) + \underline{r}_n \quad . \qquad \square$$

How the grid is chosen does not concern as long as it satisfies certain restrictions. For this chapter it is assumed that there exists a positive constant Δ independent of h such that

$$\Delta \leq \theta_n \leq 1$$

where the relative stepsize $\theta_n \equiv h_n/h$. In addition, for Section 5.1 it is required that there exists a positive constant V independent of h such that

$$\sum_{n=1}^{N} |\theta_n - \theta_{n-1}| \leq V \quad ,$$

and for Section 5.2, which discusses (k+1)-value Adams methods, it is required that changes in stepsizes do not occur more often than every k-1 steps. These restrictions on the stepsizes are weaker than those considered by Tu (1972) for the interpolatory technique of changing stepsize.

In Section 5.1 it is shown that the 1-root condition and the rth 1-order condition are sufficient for variable stepsize convergence of order r, and in Section 5.2 (k+1)-value Adams methods are shown to be kth order convergent.

Section 5.1 Multivalue Methods

For this section it is required that there exists a positive constant V independent of h such that

$$\sum_{n=1}^{N} |\theta_n - \theta_{n-1}| \le V \quad .$$

When the stepsize is varied subject to this restriction, q-convergence is not preserved unless $q = 1$. Therefore only the case $q = 1$ is considered here.

In order to show the significance of the restrictions on the stepsize, let us suppose that the stepsize is selected by means of a function $\theta(t)$:

$$t_n = t_{n-1} + h\theta(t_{n-1}) \quad .$$

(Ignore the minor problem of getting $t_N = 1$.) Then it is required that $\theta(t)$ be bounded away from zero and that $\theta(t)$ be of bounded variation.

The meaning of variable stepsize is sufficiently general to include practical schemes for choosing stepsize. Certainly any worthwhile scheme should satisfy the two requirements involving Δ and V. A more restrictive kind of stepsize selection is considered by Tu (1972), who requires that $|\theta_n - \theta_{n-1}| \le h_n\Theta'$. He shows that if h_n is controlled so that the local error estimate is equal to ε or εh_n then this requirement is met. However, for some practical schemes, θ_n does not satisfy $|\theta_n - \theta_{n-1}| \le h_n\Theta'$ for some constant Θ'. For example, the computer program DIFSUB (Gear (1971)) does not permit increases of less than ten percent in the stepsize, and so $|\theta_n - \theta_{n-1}| \ge \Delta/10$ whenever $\theta_n > \theta_{n-1}$.

[illegible]

$$[illegible]$$

[illegible]

[illegible]

$$[illegible]$$

[illegible]

[illegible]

The theory of previous chapters extends to the variable stepsize case. The definitions and theorems are restated for variable stepsize in a slightly abbreviated form. Let us adopt the abbreviated notation $||\cdot||_n \equiv ||\cdot||_{H_n}$ where $H_n = \mathrm{diag}(1, h_n, \ldots, h_n^{p-1}, \ldots, h_n^{p-1})$.

Definition of rth order convergence:

A multivalue method is convergent of order r if for any Lipschitz problem with solution $y(t) \varepsilon C^{r+p}[0,1]$,

$$||\underline{a}_n - \underline{a}(t_n; h_n)||_n = O(h^r)$$

whenever $\underline{r}_0$ is a function of h_0 satisfying

$$\underline{r}_0 = \underline{O}(h_0^{r+p-1})$$

and $\underline{r}_1 = \underline{r}_2 = \ldots = \underline{r}_N = \underline{0}$. ☐

With q = 1, it is convenient to use definitions of stability and consistency which are more conventional than the definitions of 1-stability and 1-consistency given in Chapter 3.

Definition of stability:

A multivalue method is stable if for any Lipschitz problem, there exists a constant k_0 such that

$$||\underline{a}_n-\underline{a}_n^x||_n \le k_0 \sum_{j=0}^{N} ||\underline{r}_j||_j \quad . \qquad \square$$

In order to show stability for variable stepsize, it is necessary to have $||H^{-1}S^nH||$ bounded, whereas for fixed stepsize it is sufficient to require only that $h^{p-1}||H^{-1}S^n||$ be bounded.

Definition of rth order consistency:

A multivalue method is consistent of order r if there exists a matrix $\hat{D}$ such that for any Lipschitz problem with solution $y(t)\varepsilon C^{r+p}[0,1]$, $\underline{a}^s(t;h)=\hat{D}\underline{Y}(t;h)$ satisfies

1. $||\underline{a}^s(t;h) - \underline{a}(t;h)||_H = O(h^r)$.

2. $\tilde{\underline{a}}^s(t;h) = \underline{a}^s(t;h) + \underline{O}(h^{r+p})$. $\square$

Theorem 5.1:

A multivalue method which satisfies the 1-root condition and the 1st 1-order condition is stable.

Proof: Consider a Lipschitz problem, and set $\underline{e}_n = \underline{a}_n-\underline{a}_n^x$. Then

$$\underline{e}_n = SC_n\underline{e}_{n-1} + \tilde{\underline{\ell}}\,\frac{h_n^p}{p!}\,\varepsilon_n + \underline{r}_n$$

where $\varepsilon_n = \Phi(t_n, C_n \underline{a}_{n-1}; h_n) - \Phi(t_n, C_n \underline{a}^x_{n-1}; h_n)$, and by Lemma 3.2 there exists a constant L such that

$$||\varepsilon_n|| \le L||C_n \underline{e}_{n-1}||_n \quad .$$

Write

$$\underline{e}_n = S\underline{e}_{n-1} + \underline{g}_n + \underline{r}_n \tag{5.1}$$

where

$$\underline{g}_n = S(C_n - I)\underline{e}_{n-1} + \underline{\tilde{\ell}} \frac{h_n^p}{p!} \varepsilon_n \quad .$$

Solving the difference equation (5.1) gives

$$\underline{e}_n = \sum_{j=1}^{n} S^{n-j} \underline{g}_j + \sum_{j=0}^{n} S^{n-j} \underline{r}_j \quad .$$

Hence

$$\begin{aligned} ||\underline{e}_n||_n \le & \sum_{j=1}^{n} ||H_n^{-1} S^{n-j} H_j|| \; ||\underline{g}_j||_j \\ & + \sum_{j=0}^{n} ||H_n^{-1} S^{n-j} H_j|| \; ||\underline{r}_j||_j \quad . \end{aligned} \tag{5.2}$$

By using the bound on ε_n, it follows that

$$\begin{aligned} ||\underline{g}_n||_n \le & \, (||H_n^{-1} S H_{n-1}|| \; ||C_n - I|| \\ & + \frac{h_n}{p!} ||\underline{\tilde{\ell}} h_n^{p-1}||_n \, L ||H_n^{-1} C_n H_{n-1}||) ||\underline{e}_{n-1}||_{n-1} \quad . \end{aligned}$$

Consider the (i,i)th component of $C_n - I$:

$$|(C_n - I)_{ii}| = |(\theta_n/\theta_{n-1})^i - 1|$$

$$\leq \frac{\Delta^{-i} - 1}{1 - \Delta} |\theta_n - \theta_{n-1}| \quad .$$

Since $||H_n^{-1} S H_{n-1}||$, $||\underline{\tilde{\ell}} h_n^{p-1}||_n$, and $||H_n^{-1} C_n H_{n-1}||$ are bounded, there exist constants c_0 and c_1 such that

$$||\underline{g}_n||_n \leq (c_0 |\theta_n - \theta_{n-1}| + c_1 h_n) ||\underline{e}_{n-1}||_{n-1} \quad .$$

From Lemma 3.1 we see that there is some k_1 such that $||H_n^{-1} S^{n-j} H_j|| \leq k_1$. Apply Lemma 3.3 to (5.2) with

$$w_n = ||\underline{e}_n||_n$$

$$\gamma = k_1 \sum_{j=0}^{N} ||r_j||_j$$

$$v_{n-1} = k_1 (c_0 |\theta_n - \theta_{n-1}| + c_1 h_n)$$

to get

$$||\underline{e}_n||_n \leq k_1 \sum_{j=0}^{N} ||\underline{r}_j||_j \exp\left(\sum_{j=0}^{n-1} k_1 (c_0 |\theta_{j+1} - \theta_j| + c_1 h_{j+1}) \right)$$

$$\leq k_1 \exp(k_1 c_0 V + k_1 c_1) \sum_{j=0}^{N} ||\underline{r}_j||_j$$

thus proving stability. □

With little additional effort it is possible to enlarge the class of problems for which it is known that rth order convergence is possible. The next theorem requires only that $y^{(r+p-1)}(t)$ be of bounded variation. This idea has been used by Chartres and Stepleman (1972) to analyze Euler's method.

Theorem 5.2:

Consider a multivalue method which satisfies the rth 1-order condition. Let $y(t) \varepsilon C^{r+p-1}[0,1]$ be the solution of a Lipschitz problem where $y^{(r+p-1)}(t)$ is of bounded variation on $[0,1]$. Then $\underline{a}^s(t;h) = D\underline{Y}(t;h)$ satisfies

1. $$||\underline{a}^s(t;h) - \underline{a}(t;h)||_H = O(h^r) .$$

2. $$\tilde{\underline{a}}^s(t;h) = \underline{a}^s(t;h) + \underline{O}(h^{r+p-1}v[t-h,t]) + \underline{O}(h^{r+p})$$

where $v[t-h,t]$ is the variation of $y^{(r+p-1}(t)$ on $[t-h,t]$, which shows the method is consistent of order r.

Proof: With one modification, the proof of Theorem 3.3 can also be used here. When $\underline{Y}(t;h)$ is approximated by $P\underline{Y}(t-h;h)$, the jth component of the error is

$$\binom{r+p-1}{j} \frac{h^{r+p-1}}{(r+p-1)!} (y^{(r+p-1)}(t-h)-y^{(r+p-1)}(\tau_j))$$

where $\tau_j \varepsilon (t-h,t)$ depends on t and h. Therefore

$$P\underline{Y}(t-h;h) = \underline{Y}(t;h) + \underline{O}(h^{r+p-1} v[t-h,t]) \; . \qquad \square$$

Theorem 5.3:

Consider a multivalue method which satisfies the 1-root condition and the rth 1-order condition. Let $y(t) \varepsilon C^{r+p-1}[0,1]$ be the solution of a Lipschitz problem where $y^{(r+p-1)}(t)$ is of bounded variation on [0,1]. Then the computed solution satisfies

$$||\underline{a}_n - \underline{a}(t_n;h_n)||_n \leq k_0 \sum_{n=0}^{N} ||\underline{r}_n||_n + O(h^r) \; ,$$

which shows the method is convergent of order r.

Proof: Let $\underline{a}_0^s, \underline{a}_1^s, \ldots, \underline{a}_N^s$ be the computed solution with

$$\underline{r}_0^s = C_1^{-1} \underline{a}^s(0;h_1) - \underline{a}(0;h_0)$$

$$\underline{r}_n^s = C_{n+1}^{-1} \underline{a}^s(t_n;h_{n+1}) - \underline{\tilde{a}}^s(t_n;h_n)$$

where $h_{N+1} \equiv h_N$ and $\underline{a}^s(t;h) = D\underline{Y}(t;h)$. Then

$$\underline{a}_n^s = C_{n+1}^{-1} \underline{a}^s(t_n;h_{n+1}) \; .$$

Using the triangle inequality

$$(5.3)\quad \sum_{n=0}^{N} ||\underline{r}_n^s||_n \le \sum_{n=0}^{N} ||C_{n+1}^{-1}\underline{a}^s(t_n;h_{n+1})-\underline{a}^s(t_n;h_n)||_n$$

$$+\ ||\underline{a}^s(0;h_0)-\underline{a}(0;h_0)||_0 + \sum_{n=1}^{N} ||\underline{a}^s(t_n;h_n)-\tilde{\underline{a}}^s(t_n;h_n)||_n .$$

Let us bound the terms of the first sum on the right-hand side:

$$|(C_{n+1}^{-1}\underline{a}^s(t_n;h_{n+1}) - \underline{a}^s(t_n;h_n))_i|$$

$$\le \sum_{j=0}^{r+p-1} |(\alpha_{n+1}^{-i}h_{n+1}^j - h_n^j)D_{ij} \frac{y^{(j)}(t_n)}{j!}|$$

$$\le |\theta_{n+1} - \theta_n| \sum_{j=0}^{r+p-1} \frac{\Delta^{-|j-i|_{-1}}}{1-\Delta} D_{ij} \frac{h_n^j y^{(j)}(t_n)}{j!}| \quad .$$

The first min(i,p-1)+r terms of the sum are zero; hence

$$||C_{n+1}^{-1}\underline{a}^s(t_n;h_{n+1}) - \underline{a}^s(t_n;h_n)||_n = O(h_n^r|\theta_{n+1} - \theta_n|) \ .$$

From Theorem 5.2

$$||\underline{a}^s(t_n;h_n) - \tilde{\underline{a}}^s(t_n;h_n)||_n = O(h_n^r v[t_{n-1},t_n]) + O(h_n^{r+1}).$$

Clearly

$$(5.4)\quad ||\underline{a}^s(t_n;h_n) - \underline{a}(t_n;h_n)||_n = O(h_n^r) \ .$$

Put these last three equations into (5.3):

$$\sum_{n=0}^{N} ||\underline{r}_n^s||_n = \sum_{n=0}^{N} O(h_n^r|\theta_{n+1} - \theta_n|) + O(h_0^r)$$

$$+ \sum_{n=1}^{N} (O(h_n^r v[t_{n-1},t_n]) + O(h_n^{r+1}))$$

$$= O(h^r V) + O(h^r) + O(h^r v[0,1]) + O(h^r) \ . \tag{5.5}$$

From (5.4), (5.5),and Theorem 5.1, it follows that

$$\begin{aligned}
||\underline{a}_n - \underline{a}(t_n;h_n)||_n &\le ||\underline{a}_n - \underline{a}_n^x||_n + ||\underline{a}_n^s - \underline{a}_n^x||_n + ||\underline{a}_n^s - \underline{a}(t_n;h_n)||_n \\
&\le k_0 \sum_{n=0}^{N} ||\underline{r}_n||_n + k_0 \sum_{n=0}^{N} ||\underline{r}_n^s||_n \\
&\quad + ||H_n^{-1} C_{n+1}^{-1} H_{n+1}|| \ ||\underline{a}^s(t_n;h_{n+1}) \\
&\quad - \underline{a}(t_n;h_{n+1})||_{n+1} \\
&= k_0 \sum_{n=0}^{N} ||\underline{r}_n||_n + O(h^r) + O(h_n^r)
\end{aligned}$$

thus proving the theorem. □

It has been shown that the 1-root condition and the rth 1-order condition are sufficient for convergence of order r. That they are also necessary follows from Chapter 2 because fixed stepsize methods are a special case of variable stepsize methods.

[illegible]

$$[illegible]$$

$$[illegible]$$

[illegible]

$$[illegible]$$

$$[illegible]$$

$$[illegible]$$

$$[illegible]$$

[illegible]

[illegible]

Section 5.2 Adams-Bashforth-Moulton Methods

For these methods, $p = 1$, $A = \tilde{A}$, $\ell_1 = 1$, and $\ell_2, \ell_3, \ldots, \ell_k$ are chosen so that the extraneous eigenvalues of S are all zero. Therefore we expect excellent stability properties. The relative stepsizes θ_n are required to satisfy $\Delta \le \theta_n \le 1$ for some constant $0 < \Delta \le 1$ independent of h.

Tu (1972) shows that an Adams method is stable if the value of h_n does not change more often than every k steps. Theorem 5.5 improves this result by showing that it is sufficient to require only k-1 steps between changes in stepsize.

Lemma 5.1:

Consider an A-B-M method with stability matrix S and a step selection technique with the property that $\prod_{i=1}^{n-j} SC_{n+1-i}$ is uniformly bounded for all j, n, and N such that $0 \le j \le n \le N$. Then the method is convergent of order r where $r = k$ except for the 2-value trapezoidal method ($\ell_0 = \frac{1}{2}$) in which case $r = 2$.

Proof: To prove stability, proceed as in Theorem 5.1 with a vew modifications:

Write

$$\underline{e}_n = SC_n \underline{e}_{n-1} + \underline{\ell} h_n \varepsilon_n + \underline{r}_n$$

and

$$|\varepsilon_n| \le L||C_n|| \; ||\underline{e}_{n-1}|| \ .$$

Solving the difference equation gives

$$\underline{e}_n = \sum_{j=1}^{n} \left(\prod_{i=1}^{n-j} SC_{n+1-i} \right) \underline{\ell} h_j \varepsilon_j + \sum_{j=0}^{n} \left(\prod_{i=1}^{n-i} SC_{n+1-i} \right) \underline{r}_j \ ,$$

and by the hypothesis, the matrix products are bounded by some constant.

The A-B-M methods have the property that if $y(t) \varepsilon C^{r+1}[0,1]$ is the solution of a Lipschitz problem then

$$\underline{\tilde{a}}(t;h) = \underline{a}(t;h) + \underline{O}(h^{r+1})$$

where r is defined above. The remainder of the proof is similar to the proof of Theorem 5.3. □

Theorem 5.4:

If k = 1 then the A-B-M method is convergent of order 1 for $\ell_0 \ne \frac{1}{2}$ and of order 2 for $\ell_0 = \frac{1}{2}$.

Proof:

$$\prod_{i=1}^{n-j} SC_{n+1-i} = \begin{pmatrix} 1 & (1-\ell_0)\alpha_{j+1} \\ 0 & 0 \end{pmatrix} \text{ if } n \ge j+1 \ .$$ □

Theorem 5.5:

Let $k \geq 2$, and assume that at least $k-1$ steps are taken between changes in stepsize. Then the resulting A-B-M method is convergent of order k.

Proof: S satisfies its characteristic equation $S^{k+1} - S^k = 0$. Thus

$$S(S^k \underline{\delta}_j) = S^k \underline{\delta}_j$$

and so $S^k \underline{\delta}_j$ is an eigenvector of S corresponding to 1. The eigenvalue 1 has multiplicity one and $S\underline{\delta}_0 = \underline{\delta}_0$. It follows that $S^k \underline{\delta}_j$ is a scalar multiple of $\underline{\delta}_0$ for $j = 0,1,\ldots,k$. Hence only row zero of S^k is nonzero. Thus

$$S^k = \underline{\delta}_0 \underline{\delta}_0^T S^k \quad ,$$

which implies

$$S^{k-1}(I - \underline{\ell}\, \underline{\delta}_1^T) = \underline{\delta}_0 \delta_0^T S^k \tilde{A}^{-1}$$

and

$$S^{k-1} = S^{k-1} \underline{\ell}\, \underline{\delta}_1^T + \underline{\delta}_0 \underline{\delta}_0^T S^k \tilde{A}^{-1} \quad .$$

Since $\underline{\delta}_1^T C_i S = \underline{0}^T$,

$$S^{k-1} C_i S = \underline{\delta}_0 \underline{\delta}_0^T S^k \tilde{A}^{-1} C_i S \quad . \tag{5.6}$$

The theorem follows from Lemma 5.1 if it is shown that

$$(5.7) \qquad \prod_{i=1}^{n-j} SC_{n+1-i}$$

is uniformly bounded for all j, n, and N. There are two possibilities to consider.

Case I: $n-j < 2k-2$. Since $\Delta \le \alpha_n \le \Delta^{-1}$, each of the matrices in (5.7) can be bounded, and so the product is bounded.

Case II: $n-j \ge 2k-2$. If $C_{j+2} = \ldots = C_{j+k} = I$ then define $\ell = 2$; otherwise, choose ℓ so that $2 \le \ell \le k$ and $C_{j+\ell} \ne I$. In either case it follows that

$$C_{j+\ell+1} = \ldots = C_{j+\ell+k-2} = I$$

whence the product (5.7) becomes

$$\left(\prod_{i=1}^{n-j-\ell-k-2} SC_{n+1-i} \right) S^{k-1} C_{j+\ell} S^{\ell-1} C_{j+1} \ .$$

From (5.6) we get

$$S^{k-1} C_{j+\ell} S^{\ell-1} C_{j+1} = \underline{\delta}_0 \underline{\delta}_0^T S^k \tilde{A}^{-1} C_{j+\ell} S^{\ell-1} C_{j+1} \ .$$

Notice that $SC_i \underline{\delta}_0 = \underline{\delta}_0$, and therefore

$$\prod_{i=1}^{n-j} SC_{n+1-i} = (\underline{\delta}_0\underline{\delta}_0^T) S^k \tilde{A}^{-1} C_{j+\ell} S^{\ell-1} C_{j+1} \ .$$

The right-hand side is the product of at most 2k+3 matrices each of which can be bounded. □

If there are only k-2 steps between changes in stepsize, then the method is no longer stable. For example, Tu (1972) has shown that for k = 3 the eigenvalues of $\prod_{i=1}^{n-j} SC_{n+1-i}$ are

$$1,0,0, \quad \prod_{i=j+1}^{n} \frac{\alpha_i^2(\alpha_i - 1)}{2} \quad .$$

For the sequence of stepsizes h,h/10,h,h/10,... the method is unstable because the last eigenvalue is unbounded.

Section 5.3 q-convergence and Concluding Remarks

Consider the use of a q-convergent method with fixed stepsize to integrate an appropriate Lipschitz equation. The computed value $\underline{a}_{N/2}$ is a starting value for an integration on [1/2,1], but if $q > 1$ then $i!h^{-i}(\underline{a}_{N/2})_i$ may not converge to $y^{(i)}(1/2)$ for $p-q < i \leq p-1$. Even though it seems that some vital information may be lost, the information is merely transformed, and the method

recovers the initial values at t = 1/2 by taking certain linear combinations of the components of $\underline{a}_{N/2}$. However, if a different q-convergent formula is used for the interval [1/2,1], then this information is not recovered and q-convergence is not assured on that interval.

Both formula changing and stepsize changing operate on the assumption that $\underline{a}_n$ approximates $\underline{a}(t_n)$ and not $\underline{a}^s(t_n)$. Therefore in the case of an rth order q-convergent method, these two operations generally introduce an error of order h^{r+p-q} in the computed value $\underline{a}_n$, and if $\underline{a}_n$ is regarded as a new starting value, then it is reasonable to expect nothing better than 1-convergence of order r-q+1 on the rest of the interval.

A close examination of Section 5.1 convinces one that it is difficult to prove variable stepsize q-convergence unless one assumes the 1-root condition, part 1 of the 1st 1-order condition, and part 3 of the 1st q-order condition. It is expected that there are few methods which are variable stepsize q-convergent and yet are not 1-convergent for q-differential equations. Furthermore, it is difficult to obtain necessary and sufficient conditions for variable stepsize q-convergence. Therefore it is suggested that the concept of variable stepsize q-convergence should be abandoned. That leaves us with fixed stepsize q-convergence. But fixed stepsize multivalue methods are of little practical value

because the corresponding multistep methods are more efficient. Therefore it is not worthwhile to study q-convergence with $q > 1$ for multivalue methods.

The importance of multivalue methods is due to the convenience of the scaled derivative representation. It cannot be said that multivalue formulas are a generalization of multistep formulas because it was noted in Section 2.1 that the numerical solution of a convergent corrector only method is also the solution of a stable multistep method. Thus multivalue formulas possess all of the limitations of multistep formulas. For example, Gear (1973) shows that the order of an A-stable multivalue method cannot exceed two.

The analysis of this thesis has been limited to corrector only methods and $P(EC)^M$ methods. However, the theory can be extended to include Nordsieck-type methods which use off-step points or derivatives of f. The formula for both kinds of methods can be written

$$\underline{a}_n = S\underline{a}_{n-1} + \frac{h^p}{p!}\,\underline{\Phi}(t_n,\underline{a}_{n-1};h) \ .$$

If $f = f(t)$, then there is some operator $\underline{\Omega}$, whose components are power series in hD, such that

$$\underline{\Phi}(t_n,\underline{a}_{n-1};h) = \underline{\Omega}f(t_n) \ .$$

The optimal interpretation $\underline{\Lambda}^*$ is then given by

$$\underline{\Lambda}^* = Se^{-hD}\underline{\Lambda}^* + \underline{\Omega}\,\frac{(hD)^p}{p!} \quad .$$

To show that a method is rth order consistent, it is sufficient to show that

$$||(\underline{\Lambda}^*-\underline{\Lambda})y(t)||_H = O(h^r) \quad ,$$

$$h^p||\underline{\Phi}(t,\underline{\Lambda}y(t-h);h) - \underline{\Omega}y^{(p)}(t)||_H = O(h^{r+1}) \quad .$$

And to show stability, it is enough to show that $||H^{-1}S^nH||$ is bounded and that $\underline{\Phi}$ satisfies a Lipschitz condition in its second argument.

BIBLIOGRAPHY

B. Chartres and R. Stepleman (1972) "A General Theory of Convergence for Numerical Methods," SIAM J. Numer. Anal. 9, pp.476-492.

C.W. Gear (1967) "The Numerical Integration of Ordinary Differential Equations," Math. Comp. 21, pp.146-156.

C.W. Gear (1971) Numerical Initial Value Problems in Ordinary Differential Equations, Prentice-Hall, Englewood Cliffs, N.J.

C.W. Gear (1973) "A Note on the Non-existence of Multivalue A-stable Methods of Order Greater than Two," Dept. of Comp. Sc. Report No. 569, University of Illinois, Urbana, Illinois.

P. Henrici (1962) Discrete Variable Methods for Ordinary Differential Equations, John Wiley & Sons, New York.

K.S. Miller (1968) Linear Difference Equations, W.A. Benjamin, New York.

A. Nordsieck (1962) "On the Numerical Integration of Ordinary Differential Equations," Math. Comp. 16, pp.22-49.

M.R. Osborne (1966) "On Nordsieck's Method for the Numerical Solution of Ordinary Differential Equations," BIT 6, pp.51-57.

M. Spijker (1971) "On the Structure of Error Estimates for Finite Difference Methods," Num. Math. 18, pp.73-100.

H.J. Stetter (1973) *Analysis of Discretization Methods for Ordinary Differential Equations*, Springer-Verlag, New York.

K. Tu (1972) "Stability and Convergence of General Multistep and Multivalue Methods with Variable Step Size," Dept. of Comp. Sc. Report No. 526, University of Illinois, Urbana, Illinois.